Explaining lexical diversity in dialect data: The influence of concept features

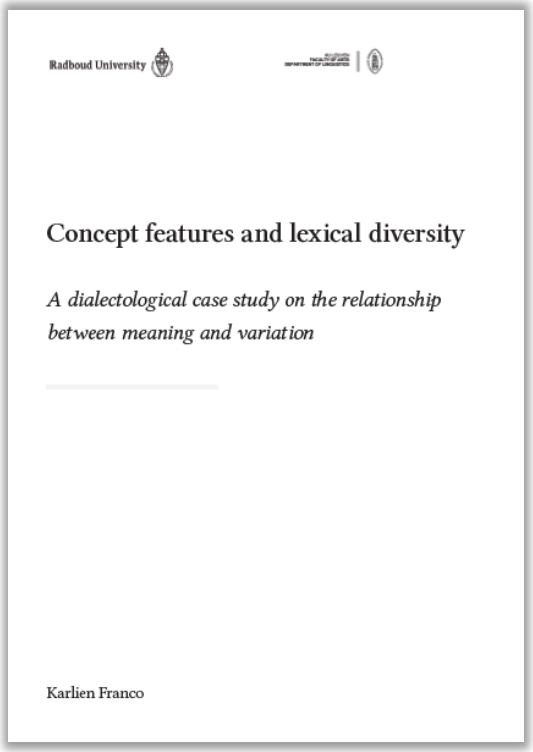
Karlien Franco

- PhD researcher at University of Leuven (QLVL)
- supervisors:

Dirk Geeraerts

Roeland van Hout

Dirk Speelman



lexical diversity across languages

= number of different words to refer to a particular concept

- variation:

  e.g. words for SNOW (cf. Boas 1911, Regier et al. 2016):

  Eskimo languages: *aput, qana, sirpoq*

  English: *snow*

  words for colors

  (e.g. Berlin & Kay 1969, Kay & Regier 2006)

  words for the senses

  (e.g. Majid & Burenhult 2014)

  ...
- stability < biological, cognitive, ... characteristics
- variation < differences in conceptualization (& lexicalization)

socio-cultural environment

categorization depends “upon the chief interests of a people; and where it is necessary to distinguish a certain phenomenon in many aspects, which in the life of the people play each an entirely independent role, many independent words may develop, while in other cases modifications of a single term may suffice”

(Boas 1911.: 26)

“language “entrenching” cognitive differences induced by cultural embodiment and cultural practice.”

(Sinha & Jensen de Lopez 2000: 37)

“The languages spoken today showcase the diverging sociocultural, environmental and linguistic histories each language has undergone.”

(Malt & Majid 2013: 592)

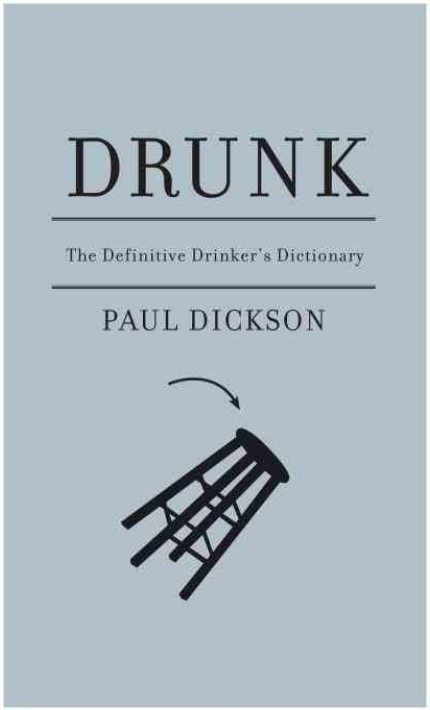
variation in lexical diversity *within one language*

e.g. SOBER:

sober, *abstinent*

DRUNK:

drunk, *blitzed*, *intoxicated*,
hammered, *pissed*, *canned*,
I'm not as think as you drunk I am, ...



Dickson (2009)

e.g. Brabantic dialects:

SLUIS 'lock (shipping)':

sluis, *sas*, ... (5 variants)

IEMAND WEERSTAAN 'to resist (someone)':

bolwerken, *volhouden*, *niet toegeven*, ... (51 variants)

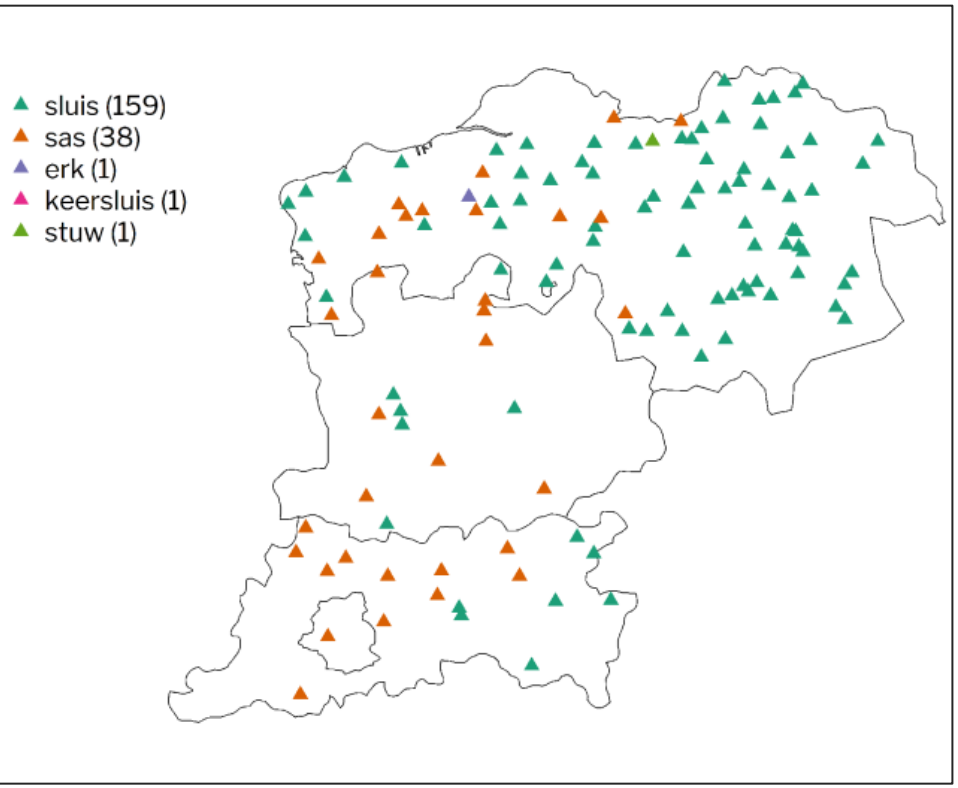
distribution of lexical items

- not only number of different variants but also how these variants are distributed in different lects
  - one word for every situation?
  - different words depending on speaker or situation?

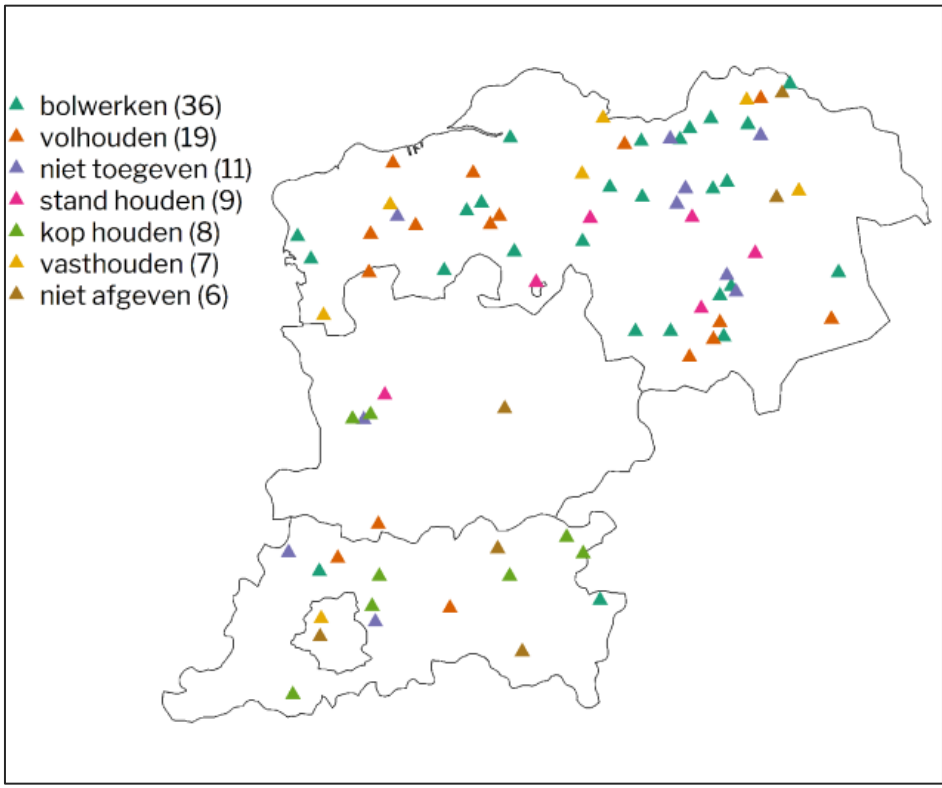
(Geeraerts, Grondelaers & Speelman 1999)

- lectal variation
  - sociolectal e.g. *blitzed* vs. *intoxicated*
  - regiolectal e.g. *pissed* (UK) vs. *canned* (US)
  - dialectal: spatial patterns of variation
  - …

taking into account (dia)lectal variation



SLUIS
'lock (shipping)'



IEMAND WEERSTAAN
'to resist (someone)'
(N > 5)

meaning influences dialectal diversity

- background
- data
- two case-studies
- conclusions

BACKGROUND

traditional explanations of variation in lexical diversity

- taboo concepts show more variation (Allan & Burridge 1988, 2006):

  → prone to rapid language change because euphemistic meaning is quickly lost

- core vocabulary is not prone to change or borrowing

  (e.g. Bochkarev et al. 2014, Hock & Joseph 1996, Tadmor et al. 2010, Thomason 2001, Zenner et al. 2014)

variation in lexical diversity in dialectology

- lack of lexical diversity ~ dialect levelling
- dialect levelling < geography and social or political factors

  e.g. mobility, population size, types of language learning, presence of state border, differences in cultural practices, ...

  (Britain 2002, 2011, Chamber & Trudgil 1980, Goossens 1964, 1972, Labov 2007, Weijnen 1967, Weinreich, Labov & Herzog 1968)

- Pickl (2013) and Szelid & Geeraerts (2008): social and semantic features cause variation in the homogeneous spatial diffusion of variants

- generally small-scale; not comprehensively researched

Cognitive Sociolinguistics

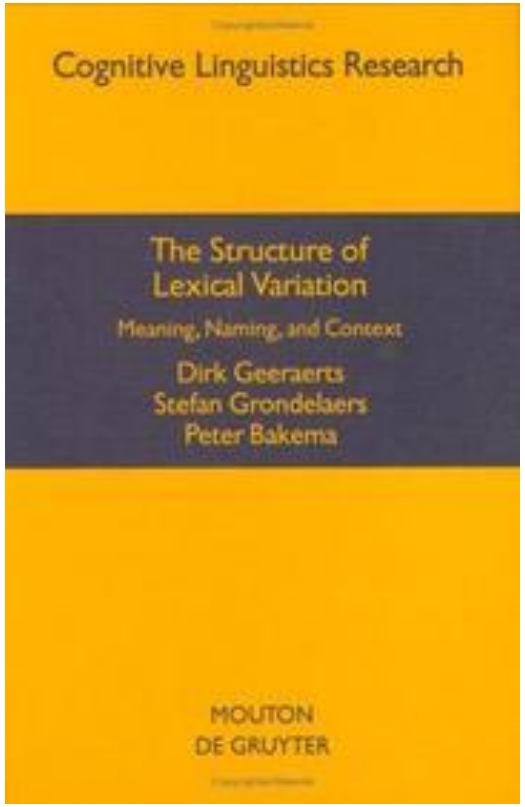
- convergence between Cognitive Linguistics and sociolinguistics

- social nature of language

- maximalist perspective on meaning
  → prototype-theoretical organization of the lexicon
  - no necessary and sufficient conditions
  - differences in degree of membership and typicality

Kristiansen & Dirven (2008)
Geeraerts, Kristiansen & Peirsman (2010)
Pütz, Robinson & Reif (2014)

two crucial distinctions



- semasiology vs. onomasiology
- formal vs. conceptual variation

Geeraerts, Grondelaers & Bakema (1994)

semasiology vs. onomasiology

semasiology ~ study of meaning

- word → meanings that word can take

  e.g. *blitzed* → DRUNK, WORN OUT

onomasiology ~ study of naming

- concept → words for the concept

  e.g. SNOW → *snow*

- onomasiological profile: all words for a concept + relative frequency (Geeraerts, Grondelaers & Speelman 1999)

  e.g. SOBER (61 mio.):*sober* 88%

  abstinent 12%

formal onomasiological variation

“jeans”

“blue jeans”

“denims”

conceptual onomasiological variation

“jeans”, “blue jeans”, “denims”

“pants”, “trousers”

focus on formal variation

"jeans", "blue jeans", "denims"

→ lexical diversity as the number of synonymous expressions & their lectal distribution

"pants", "trousers"

lexical variation in Cognitive Sociolinguistics

interacts with prototype-theoretical organization of the lexicon

- cf. Geeraerts, Grondelaers & Bakema (1994)
- e.g. typicality differences: OVERHEMD ‘dress shirt’

is more typical than

→ “*overhemd*”

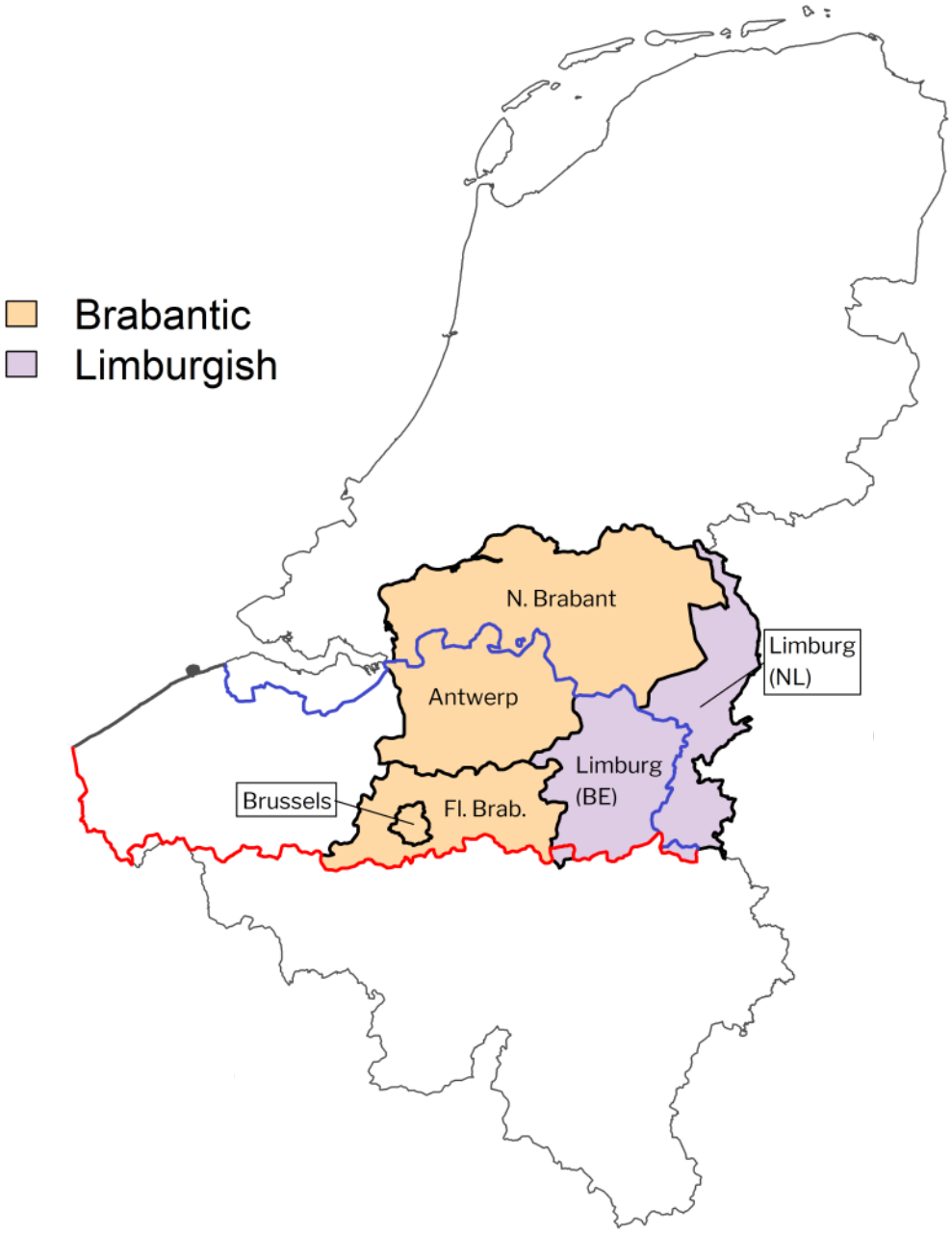
pilot studies on lexical diversity

- first to directly test correlation between semantic features and lexical diversity
  - negative affect — traditional feature
  - onomasiological vagueness — prototype theory
  - lack of onomasiological salience — prototype theory
- limited to one dialect area & one semantic field

→ are semantic concept features also influential in other semantic fields and other dialect areas?

Geeraerts & Speelman (2010), Speelman & Geeraerts (2008)

DATA


Brabantic
Limburgish
N. Brabant
Limburg (NL)
Antwerp
Limburg (BE)
Brussels
Fl. Brab.

WBD & WLD

Woordenboek van de Brabantse dialecten
‘Dictionary of the Brabantic dialects’

Woordenboek van de Limburgse dialecten
‘Dictionary of the Limburgish dialects’

WBD & WLD

- onomasiological dialect dictionaries
- digitized databases
- one or more semantic fields per case-study
- one volume = one semantic field

semantic fields (WLD)

PART 3: General vocabulary

1: Man as an individual
- The human body
- Physical activity and health
- Clothing and grooming
- Personality and feelings

2: Domestic life
- The house
- Family and sexuality
- Food and drink

3: Community life
- Society, school and education
- Celebration and entertainment
- Church and religion

4: The world versus man
- Fauna: birds
- Fauna: other animals
- Flora
- The physical and abstract world

data included in the analyses

- questionnaire data
- systematically elicited data

- data as it is available
  - no manual changes
  - advantage: large data set
  - disadvantage : little control over data

  e.g. different number of observations per concept, no specific information about background of respondents

  → aggregation

from questionnaire ...

KEUKEN-EN KOOKGEREI 2.

1. Hoe noemt u in uw dialekt:
   Het metalen voorwerp met hengsel en tuit dat dient om water in te koken?

   a. het moderne voorwerp (zie ill.) .....

2. b. het oudere voorwerp, dikwijls met ronde bodem, dat boven een haardvuur werd gehangen of in de opening van een kachel werd gezet (zie ill.)
   .....

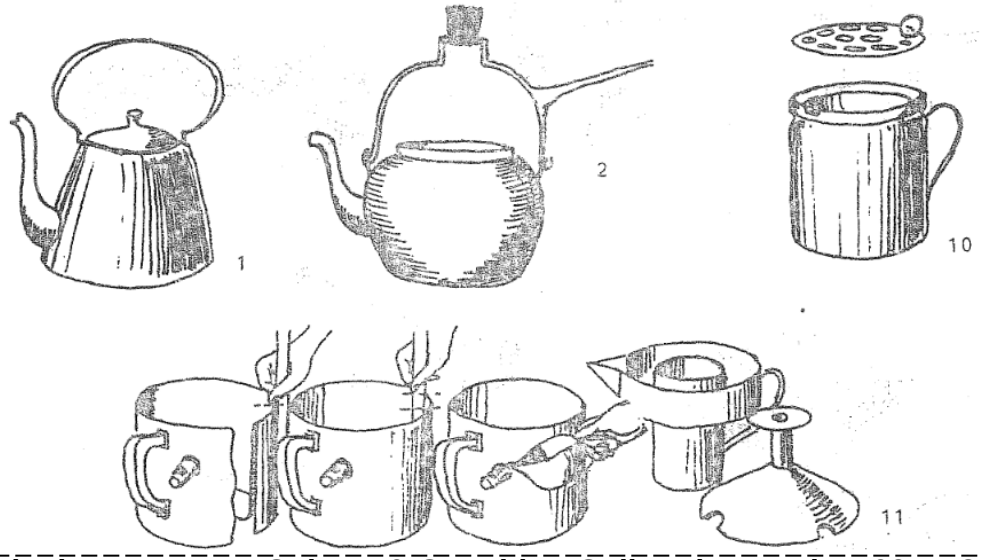
3. De pot met deksel, tuit en oor waarin koffie wordt gezet?
   .....

4. De zak waardoor het water over de gemalen koffie wordt gegoten?
   .....

5. De buikige pot met tuit en oor waarin enkel gezet?
   .....

6. Wanneer werd er thee gedronken? Als dagelijk als geneeskrachtige drank?

WOORDENBOEK VAN DE VLAAMSE DIALEKTEN : bijlage bij WVD 43 : huisraad
(Gelieve de illustraties niet terug te sturen)
Het nummer van de illustratie komt overeen met het nummer van de vraag.



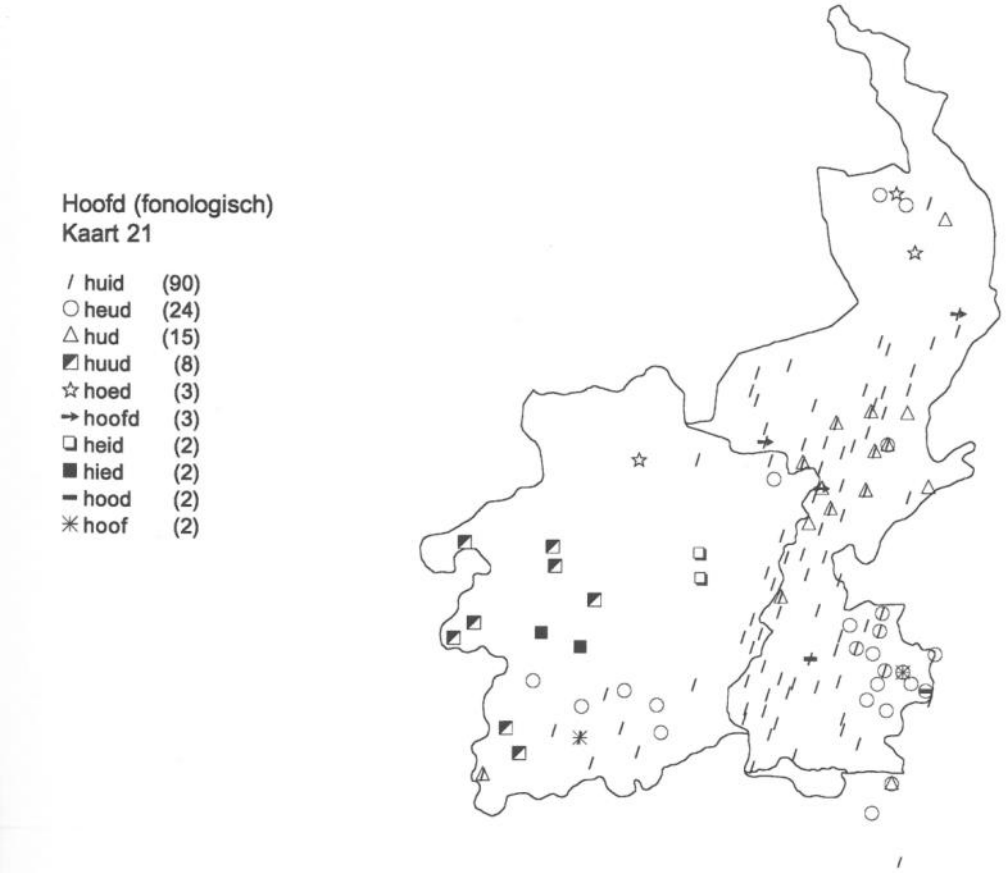
… to dataset …

concept	variant (dutchified form)	question	location	…
damesmantel ‘coat for women’	caban (fr.)	damesmantel, inventarisatie uitdrukkingen	Tervuren	…
overjas ‘overcoat’	frak	een jas die men over het colbert heen draagt	Leopoldsburg	…
…	…		…	…
vrolijk ‘cheerful’	spass (du.) haan	een opgeruimde, lichte, blijde stemming […]	Simpelveld	…
vrolijk ‘cheerful’	opgewekt	een opgeruimde, lichte, blijde stemming […]	Venlo	…
…	…		…	…

… to measurements at the level of the concept

concept	lexical diversity	predictor 1: affect sensitivity	predictor 2: vagueness	…
achterdochtig ‘suspicious’	5	sensitive	2.275	
achterhoofd ‘back of the head’	21	neutral	4.977	…
…	…	…		…
speelplaats ‘playground’	3	neutral	2.341	…
speels ‘light-hearted’	9	sensitive	3.561	…
…	…	…		…
…	…	…		…

NB: phonological variation



1.
SCALING UP

Do the semantic features have a significant and stable effect in other semantic fields and dialect areas?

Franco, Geeraerts, Speelman & Van Hout (in prep.)

pilot study	this replication study
1 semantic field the human body	6 semantic fields concrete – abstract individual – local community – society
1 dialect area Limburgish dialects	2 dialect areas Limburgish & Brabantic dialects
	$N_{concepts}$ = 3136 $N_{locations}$ = 660 N_{source} = 532 627
identical semantic features & response variable	

response variable: lexical diversity

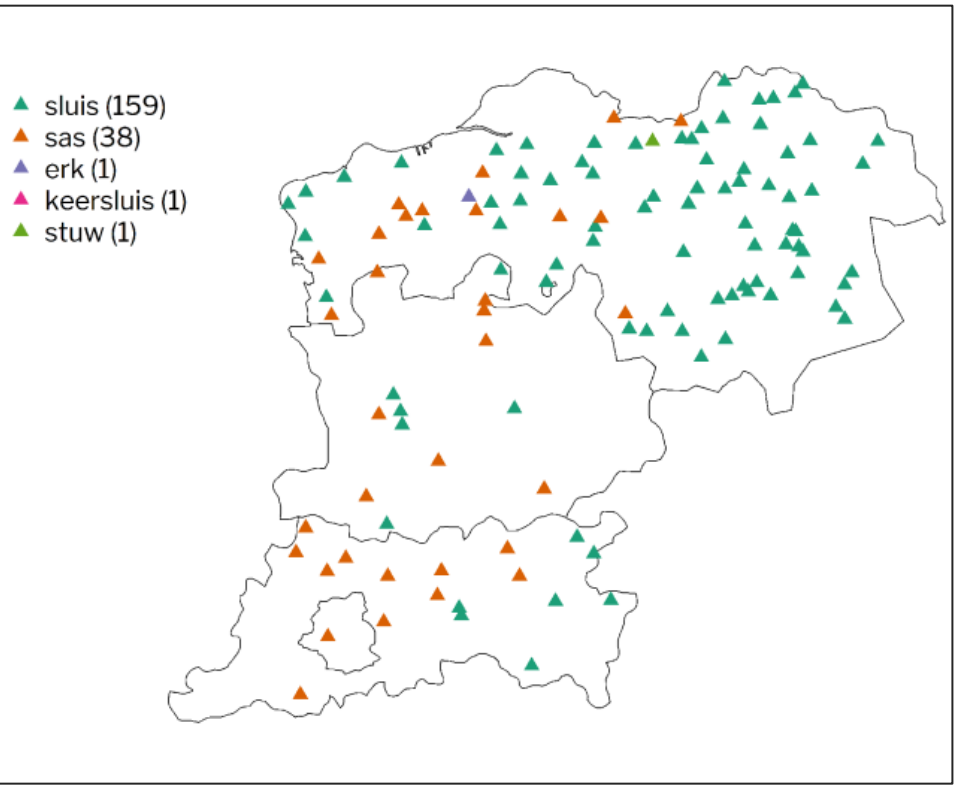
composite variable:

log(number of types * geographical fragmentation)

- number of types per concept

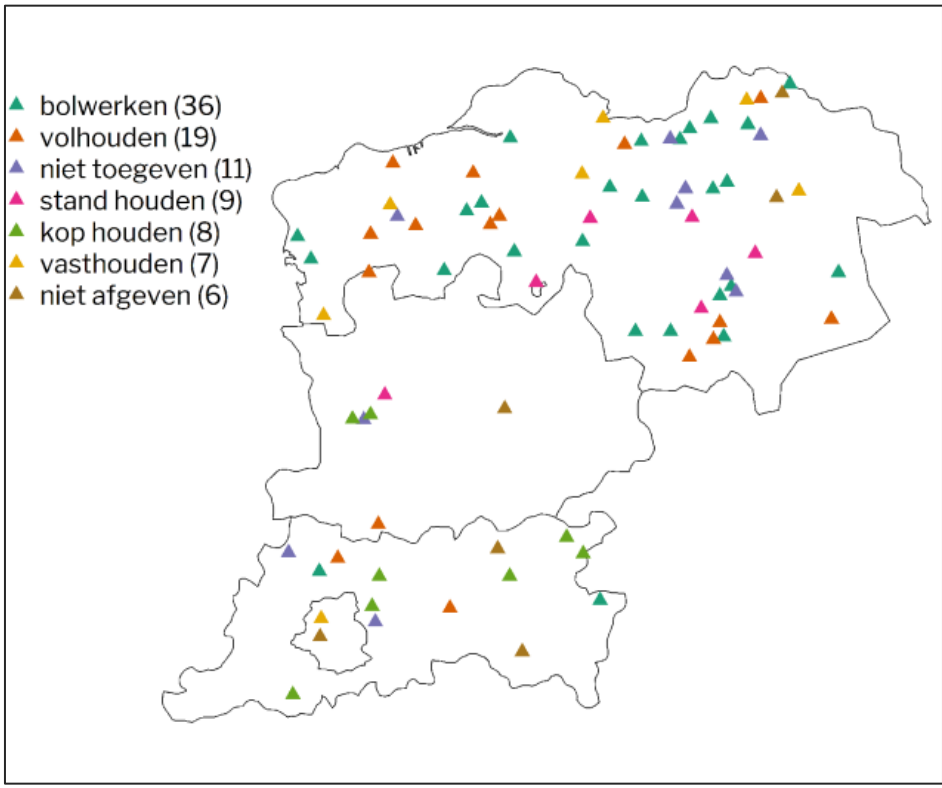
- geographical fragmentation per concept
  - lectal variation
  - quantification of spatial nature of dialectal data
  - more or less heterogeneous

less heterogeneous



SLUIS
'lock (shipping)'

more heterogeneous



IEMAND WEERSTAAN
'to resist (someone)'
(N > 5)

explanatory variables: semantic features

- onomasiological vagueness
- lack of onomasiological salience
- proneness to affect

cf. Geeraerts et al. (1994): prototype-theoretical organization of the lexicon (vagueness & salience) interacts with the structure of lexical variation

vagueness

- no necessary and sufficient conditions
- differences in degree of membership

  e.g. ‘potato’ for category ‘vegetables’

  ‘necklace’ for category ‘clothing’

- also on the onomasiological level

  → fuzziness between concepts belonging to the same semantic field

Rosch (1978, 1987)
Rosch & Mervis (1975)
Geeraerts, Grondelaers & Bakema (1994)

onomasiological vagueness

JEWELRY

CLOTHING

→ fuzziness at the edges

onomasiological vagueness

“cacti”

onomasiological vagueness

type of cactus

other type of succulent (Euphorbia)

hypothesis & operationalization

- onomasiologically vague concept ~ more lexical diversity
  < demarcation differences (Pickl 2013)
  cf. cacti: laymen vs. botanists

- operationalization: lexical non-uniqueness
  = number of lexical items that are also used for other concepts
  e.g. TO RAIN HEAVILY
  "*to rain*" → how often also used for other concepts?
  "*to pour*" (e.g. TO STORM)

onomasiological salience

how familiar is the concept for a dialect speaker?

- extension of basic-level theory
  (Berlin 1972, 1978, Berlin, Breedlove & Raven 1973, Rosch et al. 1976)

- the basic level, i.e. the generic taxon, constitutes the core of a folk-biological classification

"[a]t this rank, both plants and animals appear **perceptually most distinct** to the human classifier, and these differences in morphology and behaviour virtually **'cry out to be named'**"
(Berlin 1978: 24)

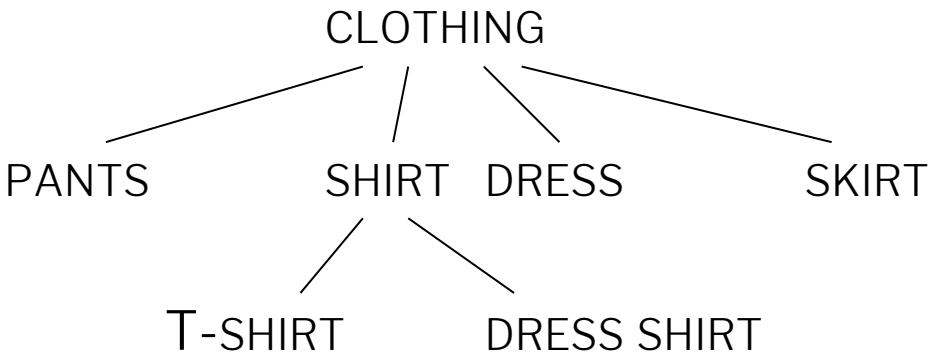
“oak”

vs. “*tree*”
vs. “*blackjack oak*”

vs. “*bird*”
vs. “*European robin*”

two problems with basic-level theory

1. assumes strict taxonomical organization



?

DAMESKLEDINGSTUK
'piece of clothing
for women'

Geeraerts et al (1994)

two problems with basic-level theory

1. assumes strict taxonomical organization

2. differences in typicality between items at the same level

e.g. OAK vs. GREVILLEA (native to Australia)

Geeraerts et al (1994)

generalized onomasiological salience

- cf. Langacker (1987) “entrenchment”
- onomasiological salience as “conceptual ‘wiring in’”:

  “a well-entrenched concept is more firmly anchored in the language user’s knowledge of the language.”

  (Geeraerts et al. 1994: 145)

- allows for differences in typicality/familiarity between items that can (but need not) be situated at the same level

- pilot studies:

  lack of onomasiological salience ~ more lexical diversity

Geeraerts et al (1994)

operationalization

- proportion of multi-word expressions
- proportion of missing places
- proportion of hapax legomena
- prevalence (Keuleers et al. 2015)
  - word-level
  - missing data

  → prevalent vs. not-prevalent/missing

affect

- taboo concepts show more lexical diversity

  cf. Allan & Burridge (1988, 2006)

  e.g. DRUNK – SOBER

- more generally:

  language users have clear positive or negative associations with words denoting particular concepts

  cf. Osgood & Tannenbaum (1957)

- pilot studies:

  negative affect ~ more lexical diversity

affect: operationalization

- forced-choice task:
  - use best judgement in deciding whether a particular concept has a connotation
  - negative, positive, neutral, unknown
- variable used in analyses: proportion of non-neutral ratings
  - certainty of non-neutrality
  - e.g. LIJKWAGEN 'hearse': 4 negative, 1 neutral rating

  → proportion of non-neutral ratings = 0.8

- only 5 respondents, but consistent with large-scale affect measurements (Moors et al 2013)
- moderate to substantial inter-rater reliability:

  Light's kappa = 0.675

testing systematicity of semantic features

- two additional variables in regression model:
  - dialect area: Limburgish / Brabantic
  - six semantic fields

- hypothesis: semantic features stable across semantic fields & across dialect areas

six semantic fields

high average degree of concreteness

- the human body

  e.g. HEAD, KNEE, FOOT, CORPULENT
- the house

  e.g. CUTLERY, TYPES OF POTS, CLEANING UP, WASHING
- celebration & entertainment

  e.g. sports & (children's) games, celebrations (e.g. CARNIVAL), the arts (e.g. SCULPTOR)

(mean concreteness: Brysbaert et al. 2014)

six semantic fields

low average degree of concreteness

- personality & feelings

  e.g. (temporary) feelings (e.g. ANGER), personality traits (e.g. TO BE SHY), behaviour (e.g. HASTY), memory & thinking (e.g. TO INFORM)
- family & sexuality

  e.g. baptism, marriage, death and burial
- society, school & education

  e.g. police, war and defence, communication, schooling, transportation

(mean concreteness: Brysbaert et al. 2014)

semantic fields along two dimensions

	concrete		abstract	
	semantic field	N	semantic field	N
individual	the human body	361	personality & feelings	703
locally-bound	the house	508	family & sexuality (WLD only)	119
societal	celebration & entertainment	471	society, school & education	974

< lay-out dictionary

RESULTS

linear regression model

- $R^2 = 0.7311$
- semantic concept features all have significant and expected effect
- no interaction effects with dialect area
- interactions with semantic field: some features have larger effect in particular semantic fields

model term	estimate	SE	p-value
intercept	2.586	0.072	< 0.001
dictionary			
WBD	0.184	0.032	< 0.001
semantic field			
the house	0.344	0.082	< 0.001
celebration & entertainment	0.059	0.079	NS
personality & feelings	0.200	0.090	< 0.05
family & sexuality	0.132	0.121	NS
society, school & education	0.274	0.072	< 0.001
lack of salience			
proportion of missing places	-1.055	0.104	< 0.001
proportion of MWE's	0.583	0.076	< 0.001
proportion of hapaxes	13.318	0.552	< 0.001
prevalence binary (missing / not prevalent)	0.228	0.032	< 0.001
vagueness			
lexical non-uniqueness	0.032	0.003	< 0.001
affect			
proportion of non-neutral ratings	0.280	0.042	< 0.001
interaction terms			
sem. field (the house) : proportion of hapaxes	1.483	0.792	< 0.1
sem. field (celebration & entertainment) : prop. of hapaxes	-3.220	0.638	< 0.001
sem. field (personality & feelings): proportion of hapaxes	-1.867	0.626	< 0.01
sem. field (family & sexuality) : proportion of hapaxes	0.736	1.205	NS
sem. field (society, school & education) : prop. of hapaxes	-1.195	0.639	< 0.1
sem. field (the house) : lexical non-uniqueness	-0.002	0.004	NS
sem. field (celebration & entertainment) : lexical non-uniq.	0.018	0.006	< 0.01
sem. field (personality & feelings): lexical non-uniqueness	-0.012	0.003	< 0.001
sem. field (family & sexuality) : lexical non-uniqueness	-0.007	0.010	NS
sem. field (society, school & education) : lexical non-uniqu.	0.001	0.003	NS
proportion of hapaxes : lexical non-uniqueness	-0.063	0.007	< 0.001

onomasiological vagueness

STUNTELEN ‘to fumble’

vague (158)

N = 288

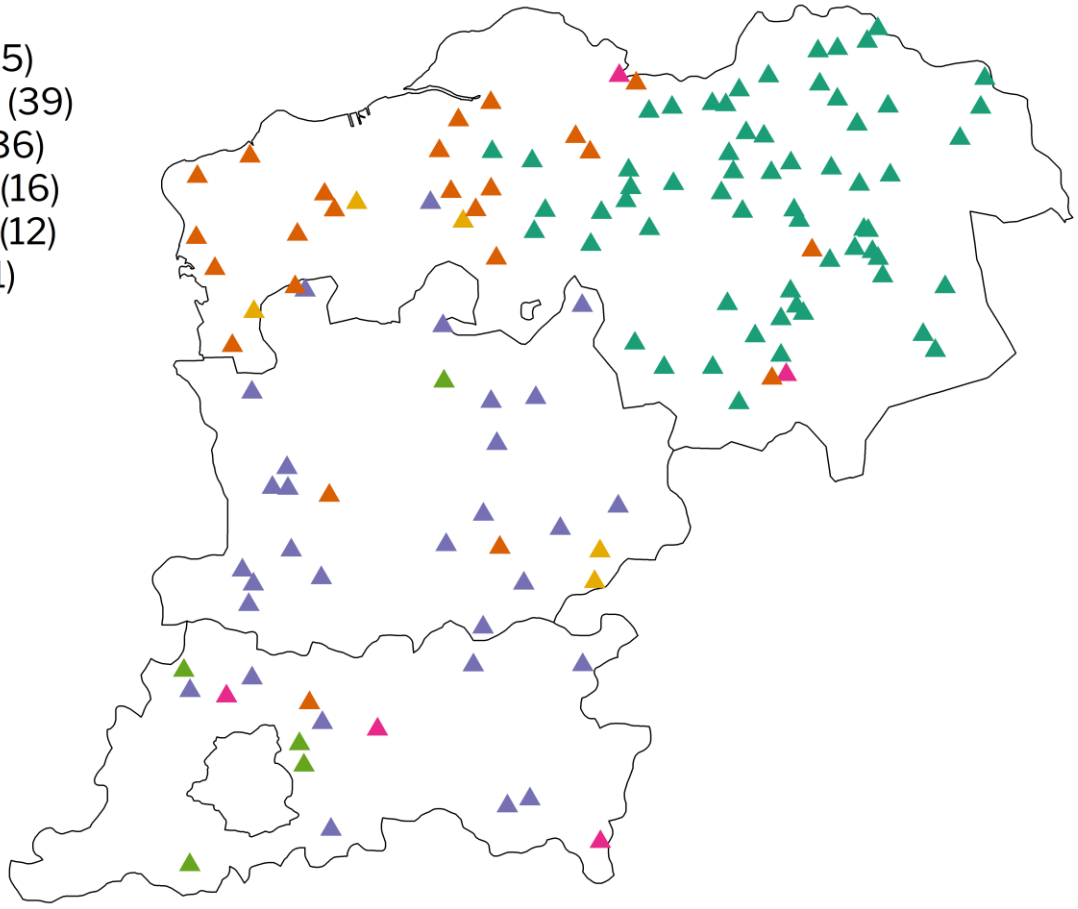
BLIND ‘blind’

not vague (0)

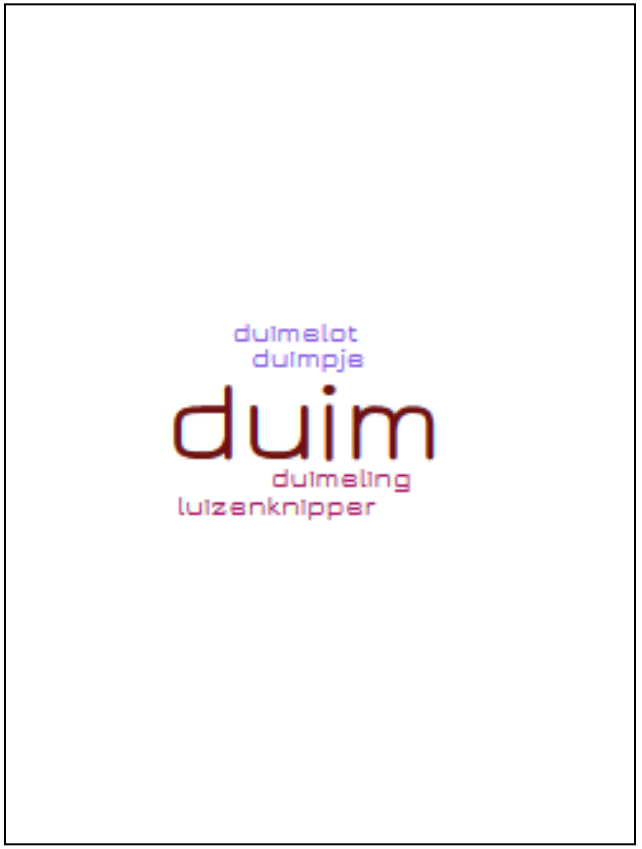
N = 327

STUNTELEN ‘to fumble’ (N > 10)

- frotten (85)
- stuntelen (39)
- prutsen (36)
- haspelen (16)
- sukkelen (12)
- prullen (11)



onomasiological salience



TENEN ‘to tiptoe’

not salient (prop. mwe = 0.876)

N = 178

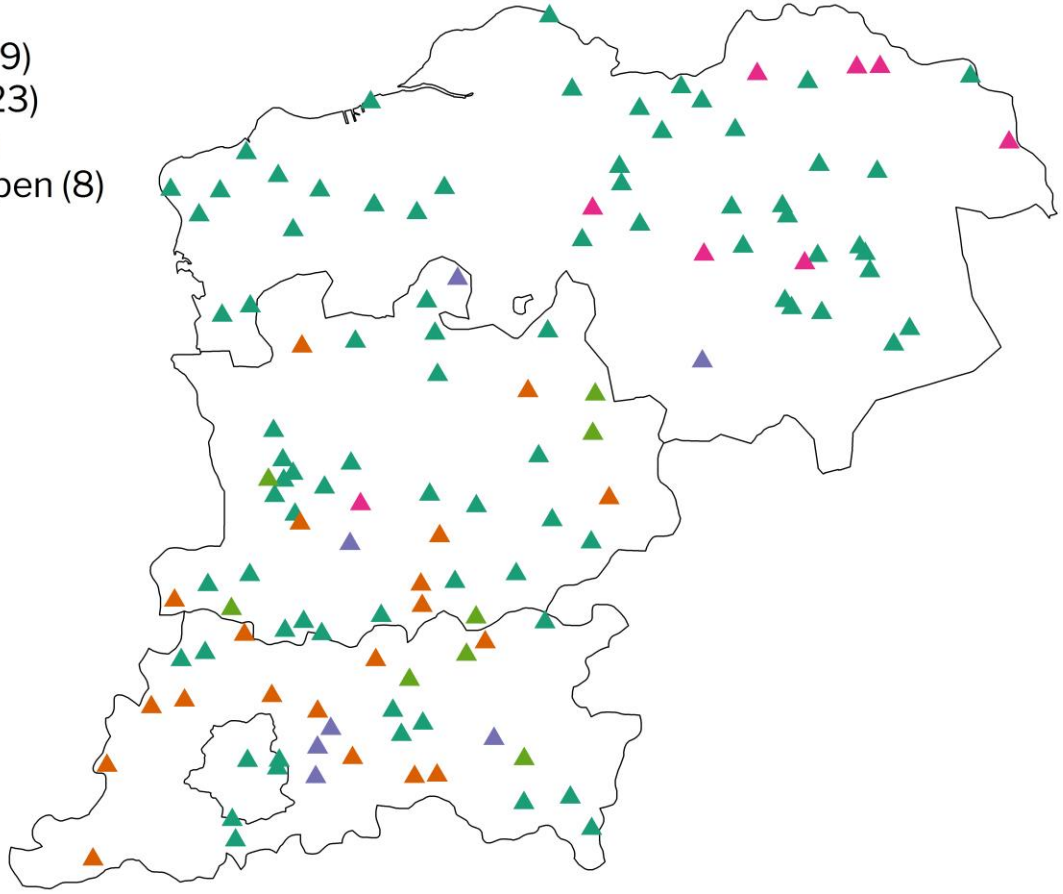
DUIM ‘thumb’

salient (prop. mwe = 0)

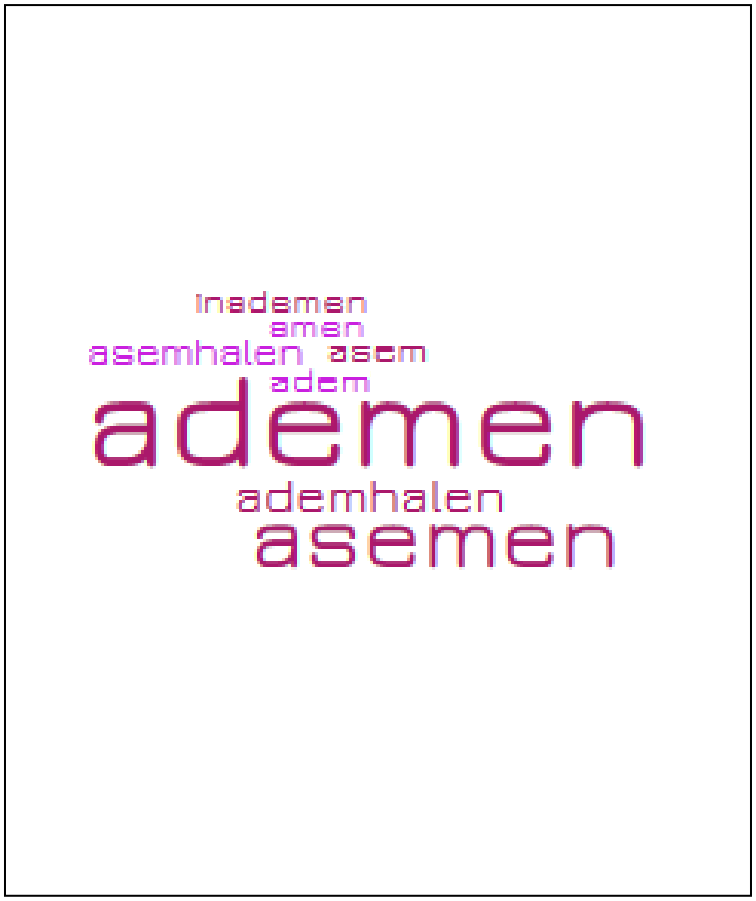
N = 209

TENEN 'TO TIPTOE' (N >5)

- op zijn tenen lopen (99)
- op zijn tippen lopen (23)
- op zijn tenen gaan (8)
- op zijn voorvoeten lopen (8)
- trippelen (8)



affect



HOUDEN VAN ‘to love’

affect-sensitive (prop not-neutral = 1)

N = 214

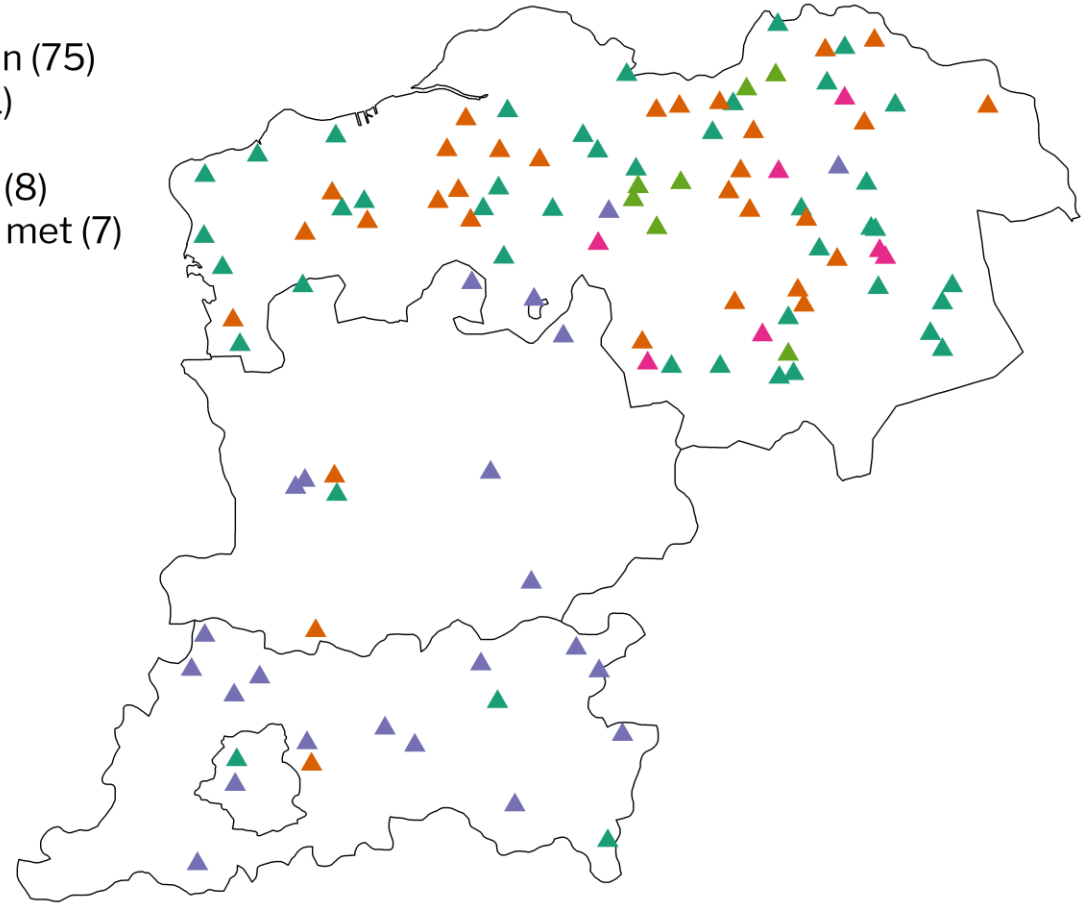
ADEMEN ‘to breathe’

neutral (prop = 0)

N = 104

HOUDEN VAN ‘to love’ (N >5)

- veel houden van (75)
- houden van (51)
- graag zien (26)
- veel houden af (8)
- veel ophebben met (7)



interim summary

- concept features influence lexical diversity
  - across semantic fields
  - across dialect areas

→ concepts that are more salient, less vague and not prone to affect show significantly less variation

- shortcoming: concept features can differ per person

  → social/lectal variation

cf. salience of cacti: laymen vs. botanists

2.
VARIATION IN CONCEPT FEATURES

Do experience-based features also correlate with lexical diversity?

Franco & Geeraerts (forthcoming)

examining variation in salience

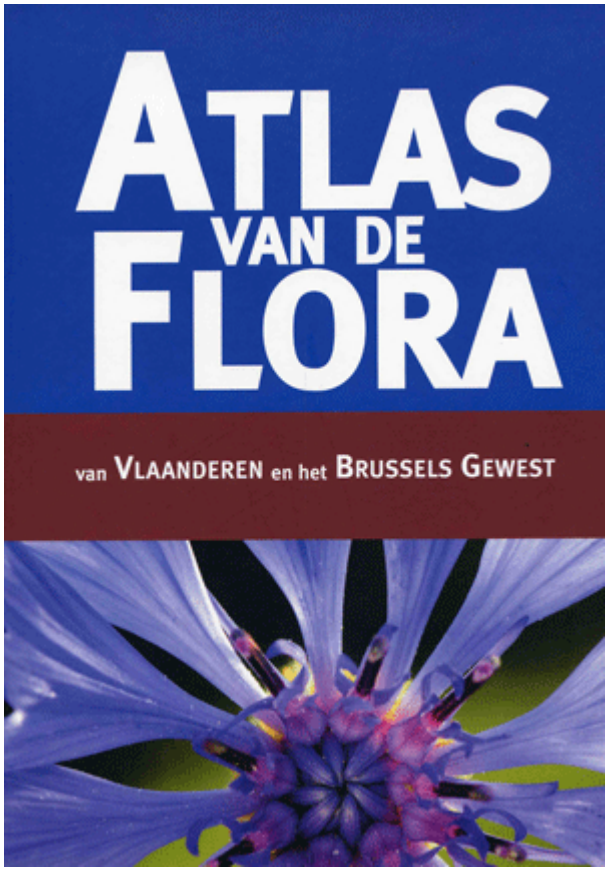
- case-study 1: salience calculated using properties of the data
  → degree of salience across the data

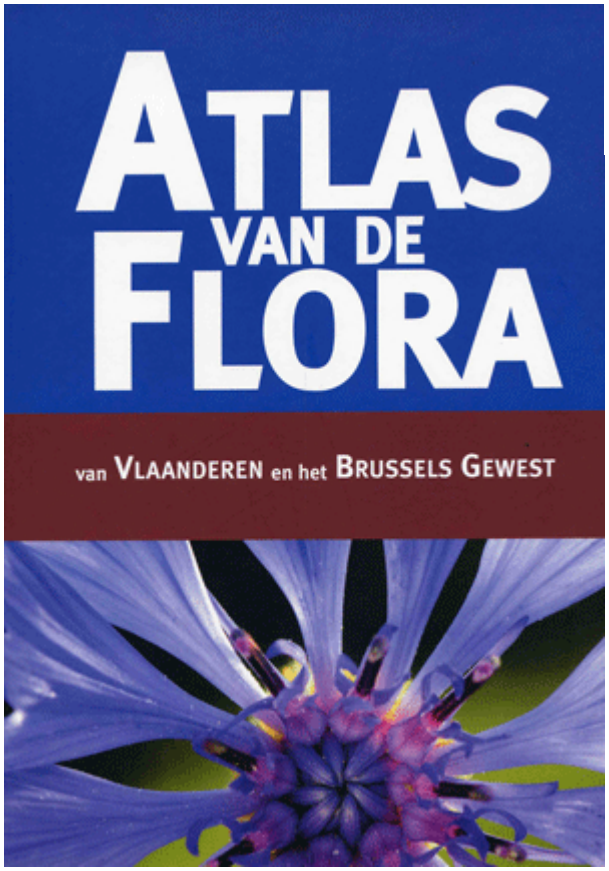
- linguistic frequency
  how often does a language user encounter a specific form to refer to the concept?
  but: geographical stratification?

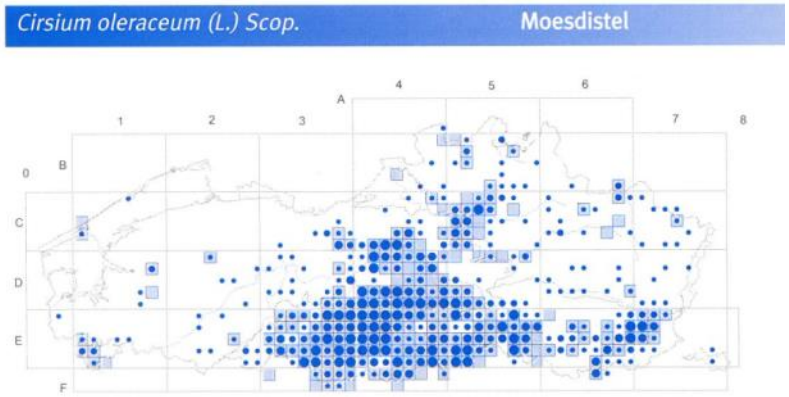
- experience with the concept/referent
  how often does a language user encounter the concept?
  → referential data to gauge (variation in) experiential salience

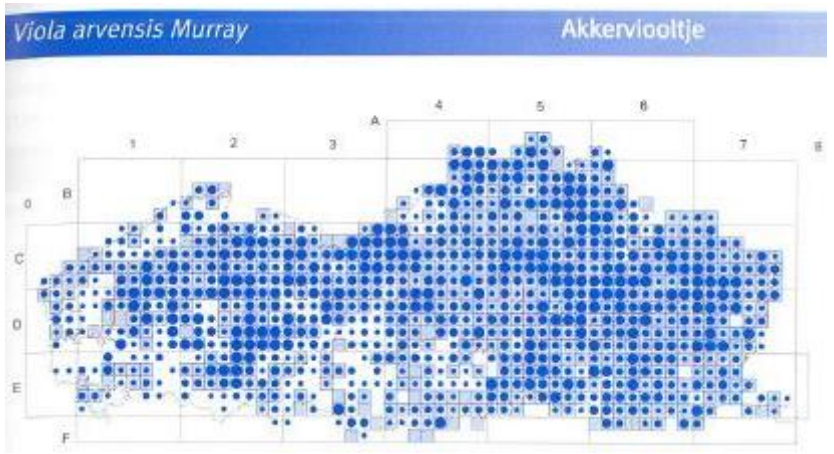
cf. Anischanka et al. (2014), Geeraerts et al. (1994)

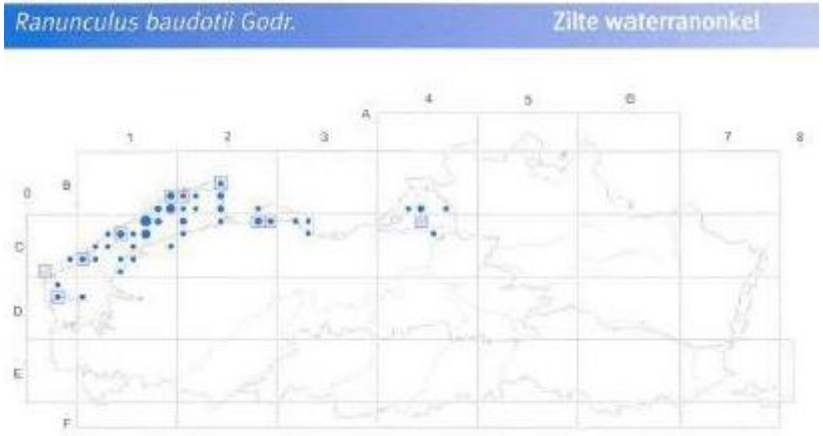
cf. Geeraerts (2016)


ATLAS
VAN DE
FLORA
van VLAANDEREN en het BRUSSELS GEWEST

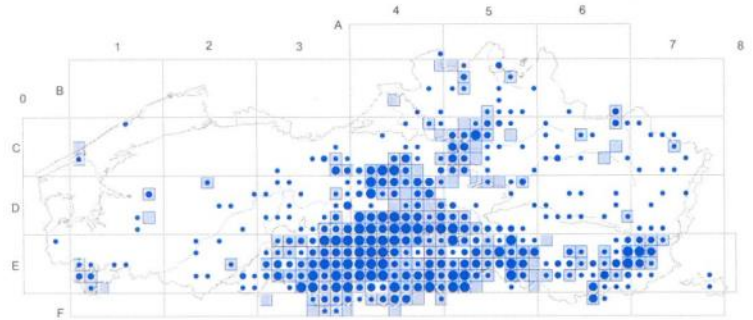

ATLAS
VAN DE
FLORA
van VLAANDEREN en het BRUSSELS GEWEST


Cirsium oleraceum (L.) Scop.
Moesdistel


Viola arvensis Murray
Akkervlooltje


Ranunculus baudotii Godr.
Zilte waterranonkel

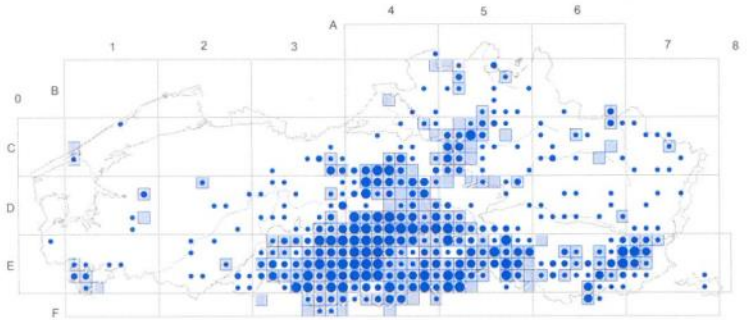
Cirsium oleraceum (L.) Scop. **Moesdistel** Paul Van den Bremt



Rode Lijst	nb
Trendindex	-0.07
KFK	7

Ecoregio	**%**
Duinen	2.4
Polder	0.4
Zand- en Zandleemstreek	12.2
Leemstreek	39.5
Kempen	7.2
Maasvallei	0

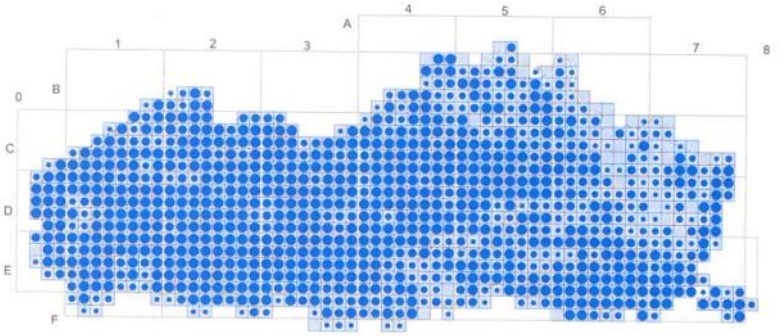
Cirsium oleraceum (L.) Scop. **Moesdistel** Paul Van den Bremt



Rode Lijst	nb
Trendindex	-0.07
KFK	7

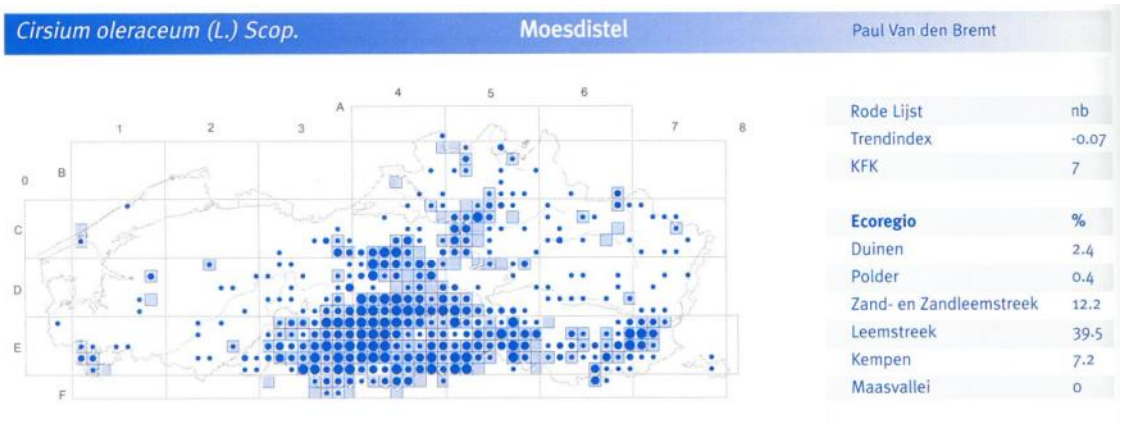
Ecoregio	**%**
Duinen	2.4
Polder	0.4
Zand- en Zandleemstreek	12.2
Leemstreek	39.5
Kempen	7.2
Maasvallei	0

Lamium album L. **Witte dovenetel** Ivan Hoste

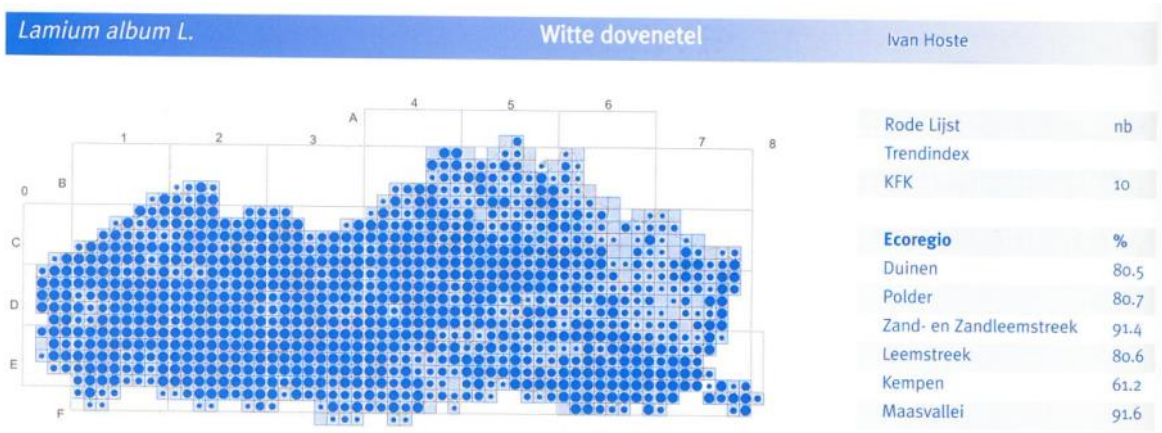


Rode Lijst	nb
Trendindex	
KFK	10

Ecoregio	**%**
Duinen	80.5
Polder	80.7
Zand- en Zandleemstreek	91.4
Leemstreek	80.6
Kempen	61.2
Maasvallei	91.6



is variation in the amount of lexical diversity per plant related to the referential frequency of a plant?

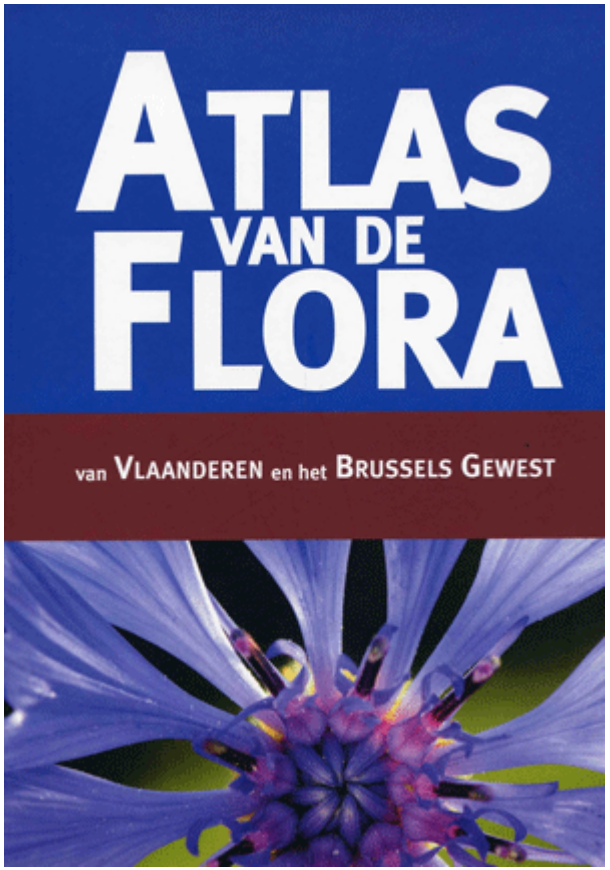


DATA & METHODS

referential data

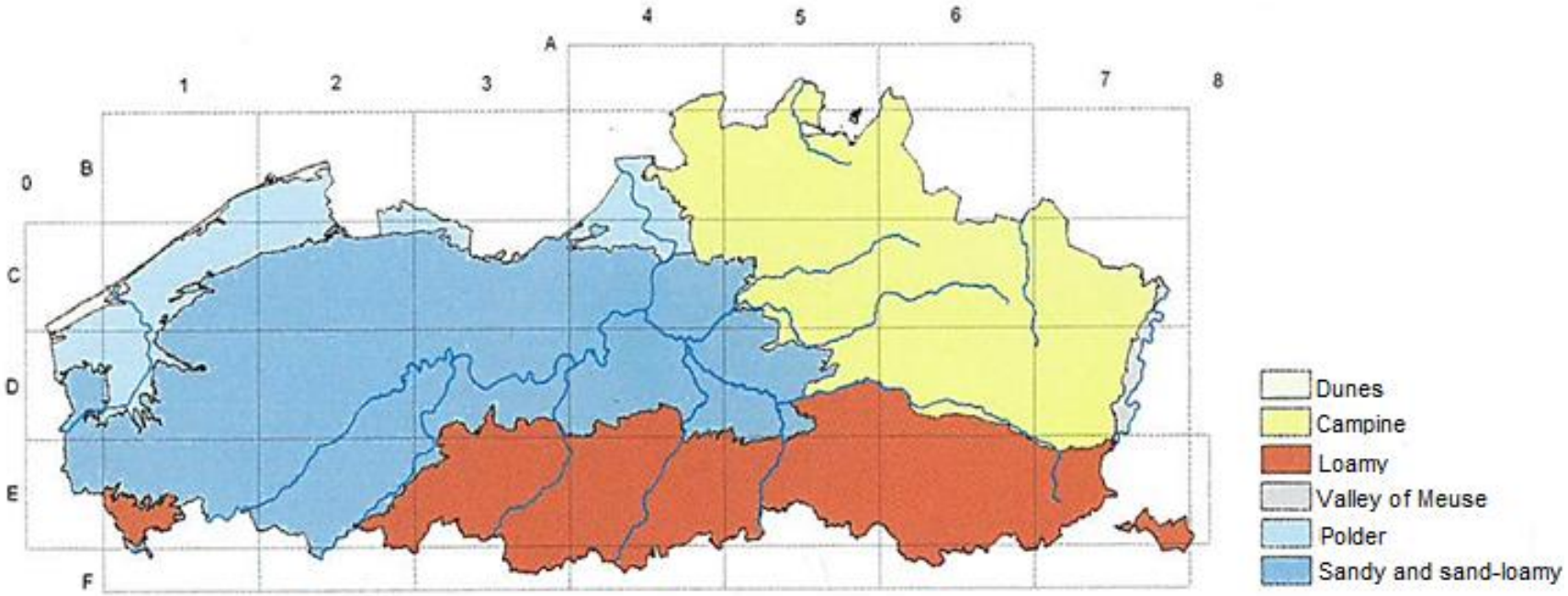
plant frequency measured on two levels:

- global frequency:
  - in Dutch-speaking Belgium as a whole
  - three measures (highly correlated)
  - absolute number of locations where plant occurs
- local frequency:
  - per ecological region
  - proportion of locations where plant occurs



Van Landuyt et al. (2006)

ecological regions

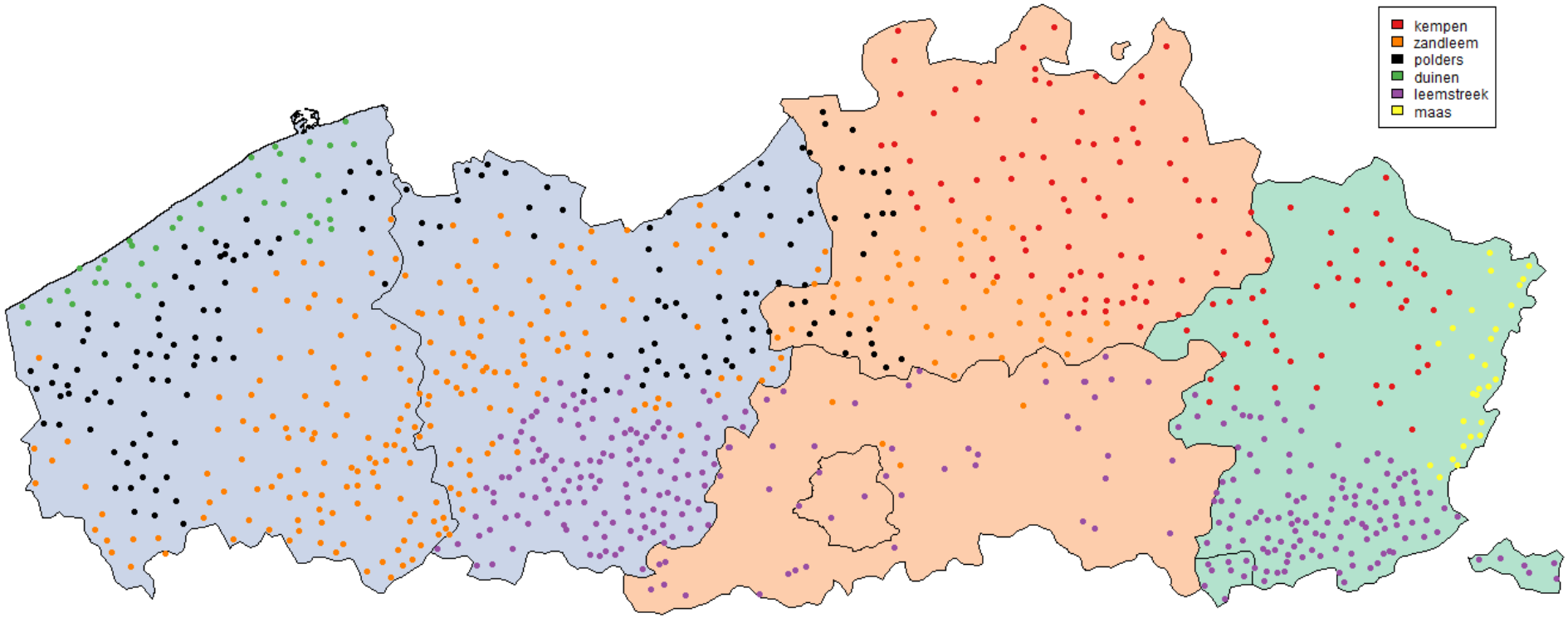


linguistic data

- semantic field of plants from WBD, WLD & WVD

(Vakgroep Nederlandse taalkunde UGent & Variaties vzw 2007)

- total number of investigated plants: N = 137



WVD WBD WLD

calculating lexical diversity

- calculated per plant per ecological region

- type-token ratio (TTR)
  - number of different types / number of different tokens
  - higher value = more lexical diversity

- internal uniformity
  - $I_Z(Y) = \sum_{i=1}^{n} F_{Z,Y}(x_i)^2$
  - relative contribution of lexical item's frequency to onomasiological profile
  - lower value = more lexical diversity

internal uniformity (I)

vergeet-mij-niet(je):
93.55% (N = 232)

blauwe kanne:
0.8% (N = 2)

onzevrouwetraantjes:
0.8% (N = 2)

...
(8 lexemes with N = 2)

$I = 0.9355^2 + 8 * (0.008^2)$
$= 0.8757$

internal uniformity (I)

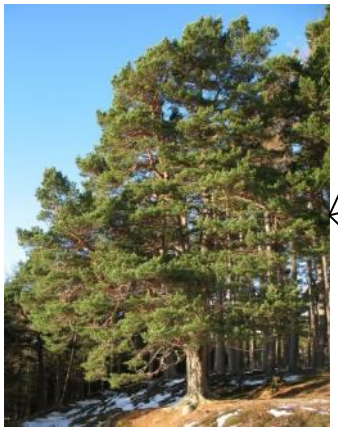
vergeet-mij-niet(je):
93.55% (N = 232)

blauwe kanne:
0.8% (N = 2)

onzevrouwetraantjes:
0.8% (N = 2)

...
(8 lexemes with N = 2)

$I = 0.9355^2 + 8 * (0.008^2)$
$= 0.8757$



den:
62.5% (N = 10)

grove den:
6.25% (N = 1)

mast:
31.25% (N = 5)

$I = 0.625^2 + 0.0625^2 + 0.3125^2$
$= 0.4922$

combining the referential and linguistic data

concept		referential frequency				linguistic diversity			
plant	ecoregion	global frequency 1	global frequency 2	global frequency 3	local frequency (%)	N tokens	N types	TTR	I
beech	Campine	2229	248	678	25.2	4	2	0.500	0.500
beech	Dunes	2229	248	678	14.6	24	3	0.125	0.462
beech	Loamy	2229	248	678	46.5	97	5	0.052	0.758
beech	Polder	2229	248	678	1.9	175	5	0.029	0.574
beech	Sand-loamy	2229	248	678	25.1	433	9	0.021	0.616

combining the referential and linguistic data

concept		referential frequency				linguistic diversity			
plant	ecoregion	global frequency 1	global frequency 2	global frequency 3	local frequency (%)	N tokens	N types	TTR	I
beech	Campine	2229	248	678	25.2	4	2	0.500	0.500
beech	Dunes	2229	248	678	14.6	24	3	0.125	0.462
beech	Loamy	2229	248	678	46.5	97	5	0.052	0.758
beech	Polder	2229	248	678	1.9	175	5	0.029	0.574
beech	Sand-loamy	2229	248	678	25.1	433	9	0.021	0.616

north of Belgium as a whole

per ecological region

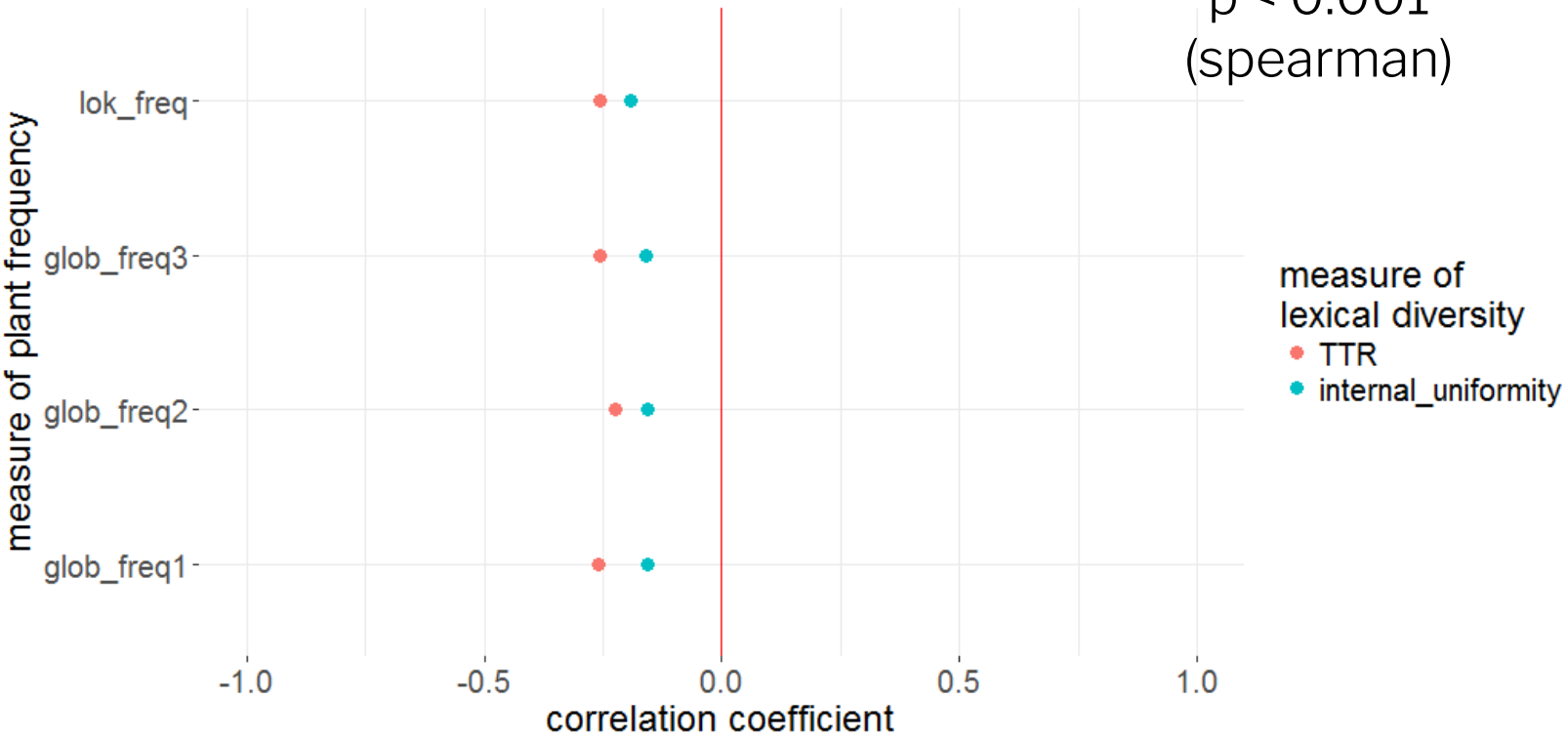
methods & expectation

negative correlation plant frequency & lexical variation:

- spearman rank correlation tests & correlation coefficients
- global vs. local frequency: linear mixed-effects regression

→ TTR: negative correlations expected
internal uniformity: positive correlations expected

RESULTS


p < 0.001
(spearman)
measure of plant frequency
lok_freq
glob_freq3
glob_freq2
glob_freq1
-1.0
-0.5
0.0
0.5
1.0
correlation coefficient
measure of
lexical diversity
TTR
internal_uniformity

plant frequency groups

		global frequency (north of Belgium)	
		frequent	infrequent
local frequency (per ecological region)	frequent	globally & locally frequent	globally infrequent
	infrequent	locally infrequent	globally & locally infrequent

plant frequency groups

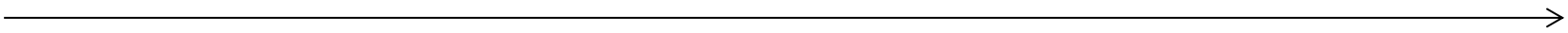
		global frequency (north of Belgium)	
		frequent	infrequent
local frequency (per ecological region)	frequent	(1) globally & locally frequent	globally infrequent
	infrequent	(2) locally infrequent	(3) globally & locally infrequent

hypothesis

(1)
globally & locally
frequent

(2)
locally infrequent

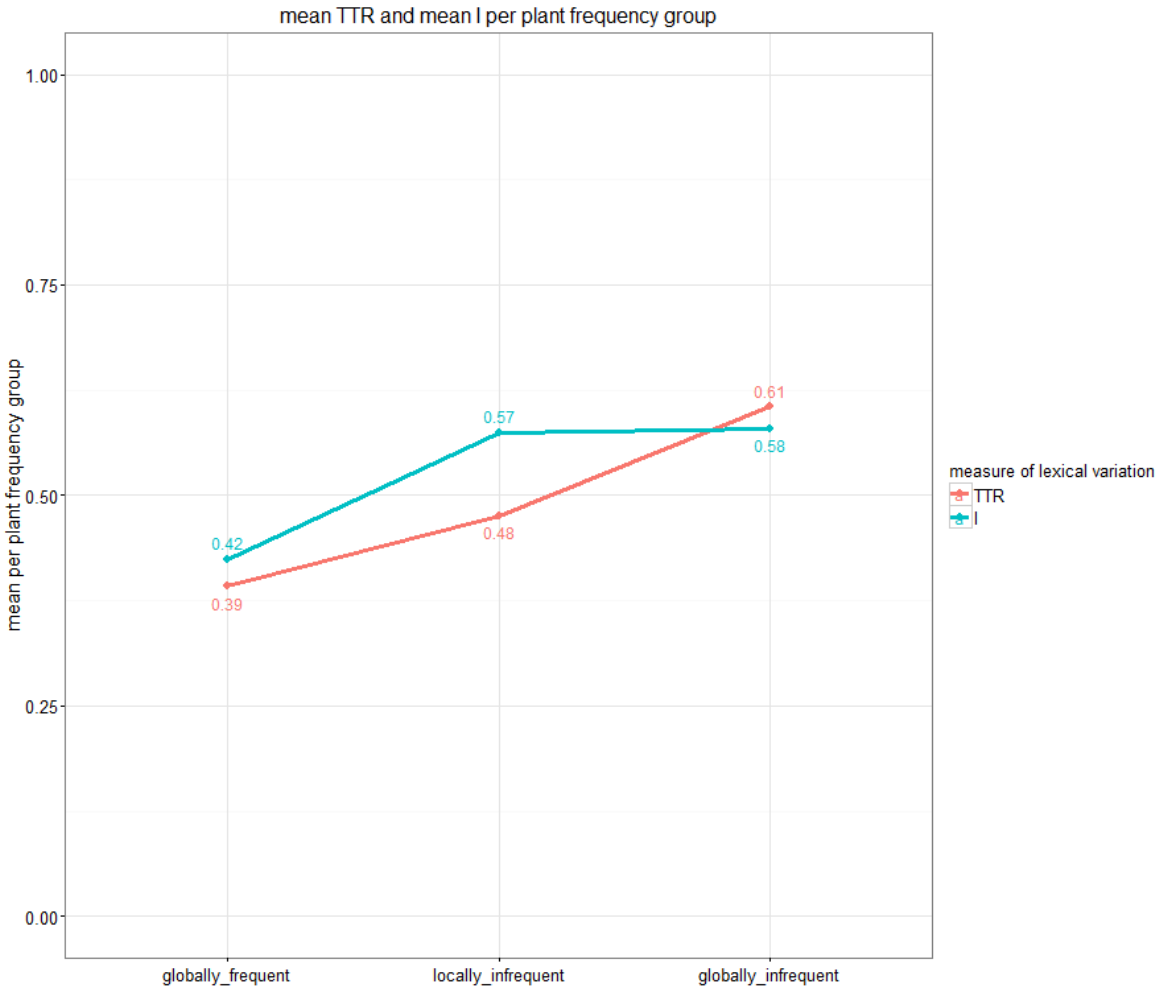
(3)
globally & locally
infrequent



less lexical diversity

more lexical diversity

global vs. local frequency



(1) (2) (3)

mean value for TTR and I per plant frequency category

expectation:

- TTR: increase
  → as expected
- I: increase
  → not as expected

stable in lmer, but small effect size

interim summary

- TTR: results as expected

  significant negative correlation between plant frequency & lexical variation

  → less frequent plants show more lexical diversity
- internal uniformity: results show opposite effect

  → names for frequent plants are not standardized enough to be picked up by I

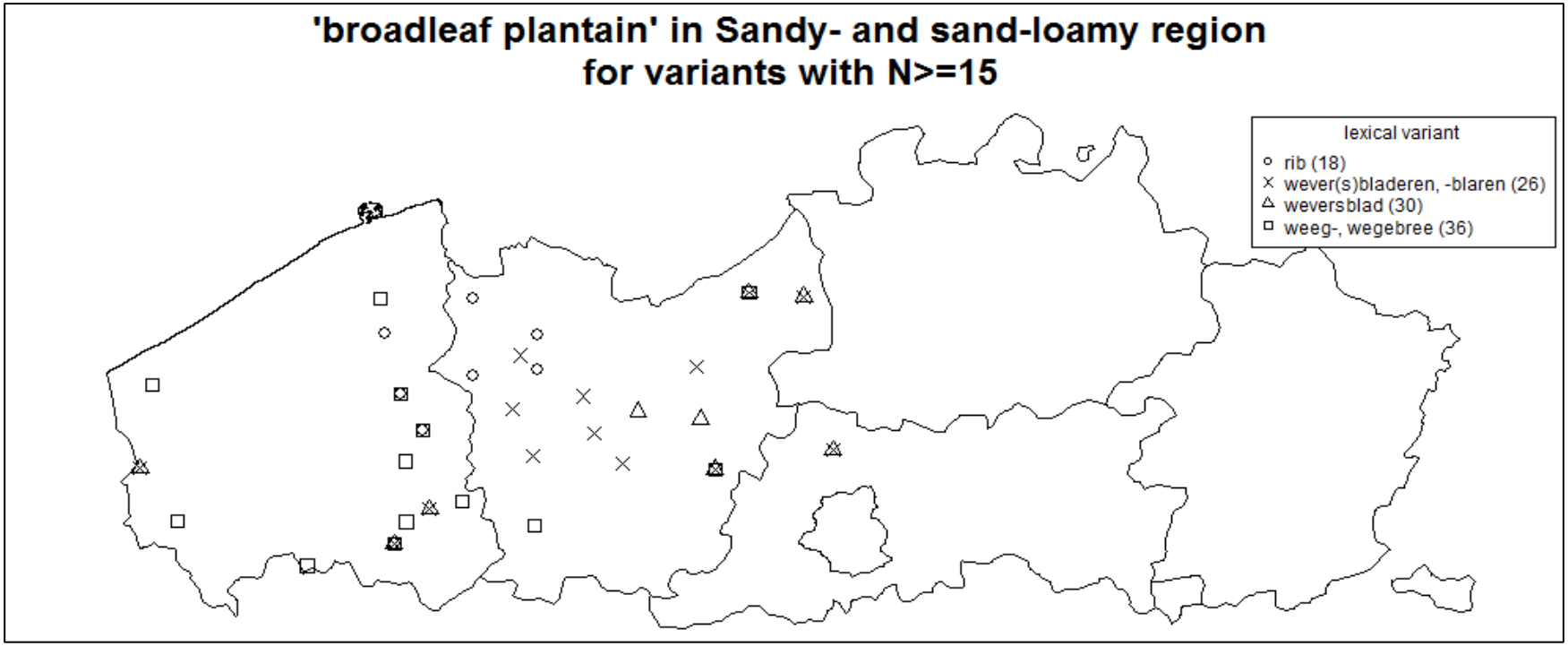
- why these diverging results?
- why small effect sizes?

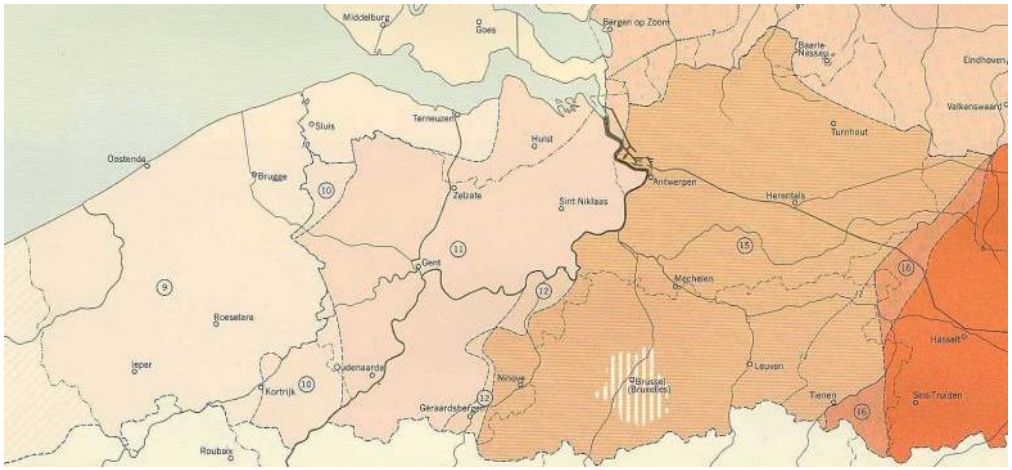
TTR vs. I

plant (ecological region)	N tokens	distribution of types	N types	TTR	I
great mullein, Loamy region	26	$lexeme_{1...18}$ occur once $lexeme_{19...22}$ occur once	22	0.846	0.050
bitter dock, Polder region	38	$lexeme_{1,2}$ occur once $lexeme_{3}$ occurs 3 times $lexeme_{4}$ occurs 4 times $lexeme_{5}$ occurs 10 times $lexeme_{6}$ occurs 19 times	6	0.158	0.338
black locust, Sandy and sand-loamy region	26	$lexeme_{1,2,3}$ occur once $lexeme_{4}$ occurs 23 times	4	0.154	0.787
forget-me-not, Dunes region	52	$lexeme_{1}$ occurs 52 times	1	0.019	1

TTR vs. I

plant (ecological region)	N tokens	distribution of types	N types	TTR	I
great mullein, Loamy region	26	$lexeme_{1...18}$ occur once $lexeme_{19...22}$ occur once	22	0.846	0.050
bitter dock, Polder region	38	$lexeme_{1,2}$ occur once $lexeme_3$ occurs 3 times $lexeme_4$ occurs 4 times $lexeme_5$ occurs 10 times $lexeme_6$ occurs 19 times	6	0.158	0.338
black locust, Sandy and sand-loamy region	26	$lexeme_{1,2,3}$ occur once $lexeme_4$ occurs 23 times	4	0.154	0.787
forget-me-not, Dunes region	52	$lexeme_1$ occurs 52 times	1	0.019	1





Daan (1969)

9: West-Flemish & Zeelandic Flemish

10: intermediate dialects between West- and East-Flemish

11: East-Flemish

15: Brabantic

other measures of experiential salience

usefulness of plant

e.g. lime tree:

- very infrequent
- little lexical diversity

→ used for artefacts

poisonousness of plant

e.g. black nightshade:

- very frequent
- a lot of lexical variation

→ negative connotation

etc.

other measures of experiential salience

the more experientially salient the plant, the less lexical diversity

→ less diversity for plants that...

are useful

higher edibility rating (pfaf.org)

higher medicinal rating (pfaf.org)

are poisonous (vs. not poisonous; data U Cornell)

results: TTR in WVD

- edible plants show significantly less lexical diversity ($p < 0.01$, Adj R^2: 0.065)
- plants that are useful for medicinal applications show significantly less lexical diversity ($p < 0.05$, Adj R^2: 0.039)
- poisonousness: NS, but on average, poisonous plants show more lexical diversity
- similar trends in WLD & WBD, but NS

discussion

experiential salience correlates negatively with lexical diversity
(referential frequency, usefulness, poisonousness etc.)

experiential salience alone cannot account for variation
cf. small effect sizes

→ communicative relevance

CONCLUSIONS

conclusions

- lexical diversity occurs across languages, but also within one language

- dialectal lexical diversity correlates with concept features
  - across dialect areas and across semantic fields
  - concept features can be prone to lectal variation

- the amount of lexical diversity is higher for:
  - onomasiologically vague concepts
  - experientially and onomasiologically less salient concepts
  - concepts prone to affect

future research

- how does the lectal dimension interact with conceptual variation across languages?

- to what extent are these features relevant across language?

  same effect? vagueness & affect vs. salience?

- to what extent can these features account for diachronic variation?

  e.g. is the speed of lexical change slower for more salient concepts? (cf. Bochkarev et al. 2014)

Thank you!

Questions? Suggestions?

karlien.franco@kuleuven.be
www.karlienfranco.com

REFERENCES

Allan, K. & Burridge, K. (1988). Euphemism, dysphemism, and cross-varietal synonymy. *La Trobe working papers in linguistics*, *1*, 1-16.

Allan, K. & Burridge, K. (2006). *Forbidden words: Taboo and the censoring of language*. Cambridge: Cambridge University Press.

Berlin, B. (1972). Speculations on the growth of ethnobotanical nomenclature. *Language in Society, 1*(1), 51-86.

Berlin, B. (1978). Ethnobiological classification. In Rosch, E. & Lloyd, B. *Cognition and categorization*. New York: Wiley. 9-26.

Berlin, B. & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California press.

Berlin, B., Breedlove, D. & Raven, P. (1973). General Principles of Classification and Nomenclature in Folk Biology. *American Anthropologist, 75*(1), 214-242.

Boas, F. (1911). Introduction. In Boas, F. *Handbook of American Indian languages. Part 1* (Smithsonian institution. Bureau of American ethnology. Bulletin 40) Washington: Government Printing Office. 1-83.

Bochkarev V., Solovyev V., Wichmann S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface*, *11*: 20140841.

Britain, D. (2002). Space and spatial diffusion. In Chambers, J.K., Trudgill, P. & Schilling-Estes, N. *The Handbook of Language Variation and Change*. Malden/Oxford: Blackwell. 603-637.

Britain, D. (2011). The heterogeneous homogenisation of dialects in England. *Taal & Tongval 63*(1), 43-60.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W. & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150*, 80-84.

Chambers, J. K. & Trudgill, P. (1998). *Dialectology* (Cambridge textbooks in linguistics). Cambridge: Cambridge University press.

Daan, J. (1969). Dialecten. In Daan, J. & Blok, D.P. *Van randstad tot landrand. Toelichting bij de kaart: Dialecten en naamkunde* (Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse akademie van wetenschappen te Amsterdam 37). Amsterdam: Noord-Hollandsche uitgeversmaatschappij. 7-43.

Dickson, P. (2009). *Drunk: The Definitive Drinker's Dictionary*. Melville House.

Franco K., Geeraerts D. (forthcoming). Botany meets lexicology: the relationship between experiential salience and lexical diversity. In: Fon J. (Eds.), *Dimensions of Diffusion and Diversity.* Berlin: De Gruyter Mouton.

Franco, K., Geeraerts, D., Speelman, D. & Van Hout, R. (in prep). Concept characteristics and the degree of lexical diversity in two Dutch dialect areas. To be submitted to *Cognitive Linguistics*.

Geeraerts, D. (2016). Entrenchment as onomasiological salience. In Schmid, H. (ed.). *Entrenchment and the psychology of language learning. How we reorganize and adapt linguistic knowledge*. Berlin, Boston: De Gruyter Mouton. 153-174.

Geeraerts, D. & Speelman, D. (2010). Heterodox concept features and onomasiological heterogeneity in dialects. In Geeraerts, D., Kristiansen, G. & Peirsman, Y. (eds.). *Advances in cognitive sociolinguistics* (Cognitive linguistics research 45). Berlin: De Gruyter Mouton. 23-39.

Geeraerts, D., Grondelaers, S. & Bakema, P. (1994). *The structure of lexical variation: Meaning, naming, and context* (Cognitive linguistics research 5). Berlin: De Gruyter Mouton.

Geeraerts, D., Grondelaers, S. & Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.

Geeraerts, D., Kristiansen, G. & Peirsman, Y. (2010). *Advances in Cognitive Sociolinguistics* (Cognitive linguistics research 45). Berlin: De Gruyter Mouton.

Goossens, J. (1964). *Enkel- en veeltoepasselijkheid van betekenaars op de taalkaart. In Taalgeografie en semantiek. Lezingen gehouden voor de dialectencommissie der koninklijke Nederlandse academie van wetenschappen op 27 december 1962 door dr. J. Goossens en dr. Jan van Bakel (*Bijdragen en mededelingen der dialectencommissie van de koninklijke Nederlandse akademie van wetenschappen ter Amsterdam XXVIII). Amsterdam: Noord-Hollandse uitgevers maatschappij. 3-27.

Goossens, J. (1972). *Inleiding tot de Nederlandse Dialectologie* (Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie 44). Tongeren: Michiels.

Hock, H.H. & Joseph, B.D. (1996). *Language history, language change, and language relationship: An introduction to historical and comparative linguistics* (Trends in linguistics. Studies and monographs 93). Berlin: De Gruyter.

Kay, P. & Regier, T. (2006). Language, thought and color: recent developments. *TRENDS in Cognitive Sciences 10*(2), 51-54.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1-62.

Kristiansen, G. & Dirven, R. (2008). *Cognitive Sociolinguistics* (Cognitive Linguistics Research 39). Berlin/New York: De Gruyter Mouton.

Labov, W. (2010). *Principles of linguistic change. Volume 3: Cognitive and cultural factors* (Language in society 39). Chichester: Wiley-Blackwell.

Langacker, R. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Stanford: Stanford University Press.

Majid, A. & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition, 130*(2), 266-270.

Malt, B.C. & Majid, A. (2013). How thought is mapped into words. *WIREs Cogn Sci*, *4*, 583–597.

Malt, B.C. & Majid, A. (2013). How thought is mapped into words. *WIREs Cogn Sci*, *4*, 583–597.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A., De Schryver, M., De Winne, J. & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4300 Dutch words. *Behavior Research Methods, 45*(1), 169-177.

Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana: University of Illinois press.

Pickl, S. (2013). Lexical meaning and spatial distribution. Evidence from geostatistical dialectometry. *Literary and Linguistic Computing*, *28*(1), 63 -81.

Pütz, M., Robinson, J.A. & Reif, M. (2014). *Cognitive sociolinguistics: Social and cultural variation in cognition and language use* (Benjamins current topics 59). Amsterdam: John Benjamins Publishing Company.

Regier T, Carstensen A, Kemp C (2016) Languages Support Efficient Communication about the Environment: Words for Snow Revisited. *PLoS ONE 11*(4): e0151138.

Rosch, E. (1978). Principles of Categorization. In Rosch, E. & Lloyd, B.B. (eds.), *Cognition and Categorization.* New York: Wiley. 27–48.

Rosch, E. (1987). Linguistic relativity. *ETC: A Review of General Semantics*, 44(3), 254-279.

Rosch, E. & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605

Sinha, C. & Jensen de López, K. (2001). Language, Culture and the Embodiment of Spatial Cognition. *Cognitive Linguistics, 11*(1-2), 17-41.

Speelman, D. & Geeraerts, D. (2008). The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing, 2*(1-2), 221-242.

Szelid, V. & Geeraerts, D. (2008). Usage-based dialectology. Emotion concepts in the Southern Csango dialect. *Annual Review of Cognitive Linguistics, 6*, 23 - 49.

Tadmor, U., Haspelmath, M & Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, *27*(2), 226-246.

Thomason, S. (2001). *Language contact*. Edinburgh: Edinburgh university press.

Van Landuyt, W., Hoste, I., Vanhecke, L., Van den Bremt, P., Vercruysse, W. & De Beer, D. (2006). *Atlas van de Flora van Vlaanderen en het Brussels Gewest*. Brussels: Research Institute for Nature and Forest, National Botanic Garden of Belgium & Flo.Wer.

WBD = *Woordenboek van de Brabantse Dialecten*. (1967-2005). Assen: Van Gorcum / Amsterdam: Gopher.

Weijnen, A. (1966). *Nederlandse dialectkunde* (Studia theodisca 10). Assen: Van Gorcum.

Weinreich, U., Labov, W. & Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W., Malkiel, Y. (eds.). *Directions for Historical Linguistics*. Austin: University of Texas Press. 95-195.

WLD = *Woordenboek van de Limburgse Dialecten*. (1983-2008). Assen: Van Gorcum / Amsterdam: Gopher.

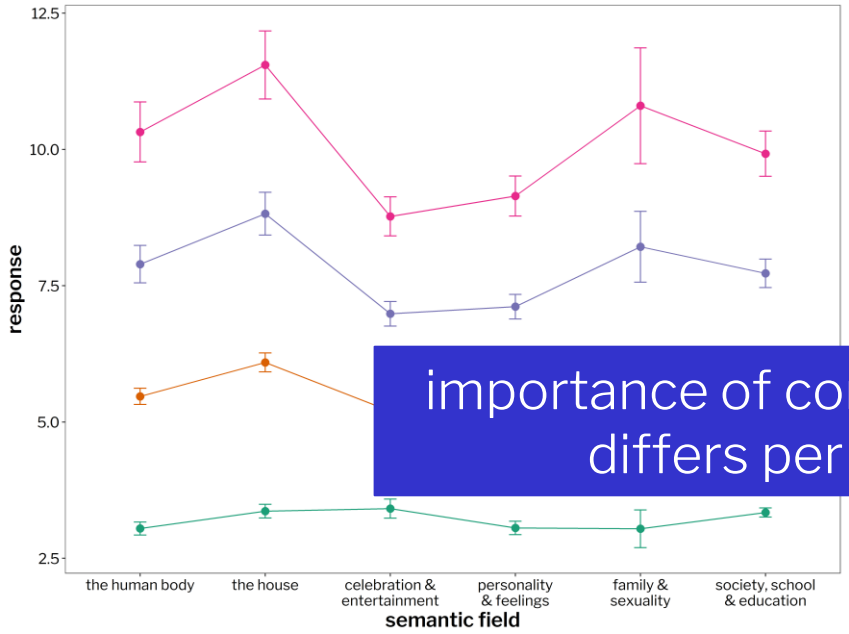
WVD = *Woordenboek van de Vlaamse Dialecten*. (1972-). Gent: Academia Press.

Zenner, E., Speelman, D. & Geeraerts, D. (2014). Core vocabulary, borrowability, and entrenchment: A usage-based onomasiological approach. *Diachronica, 31*(1), 74-105.

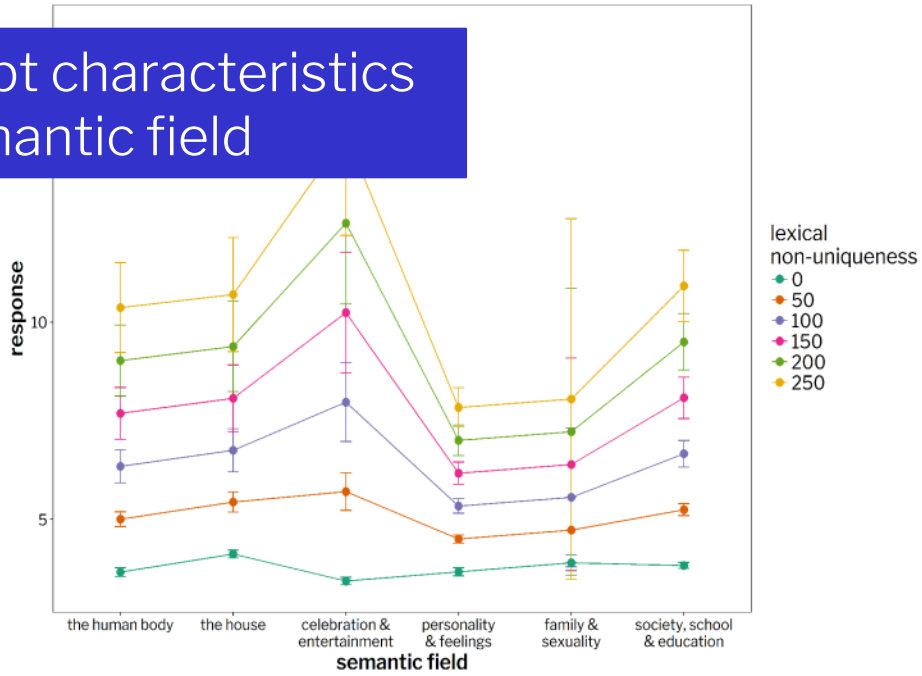
EXTRA

INTERACTIONS MODEL CASE-STUDY 1

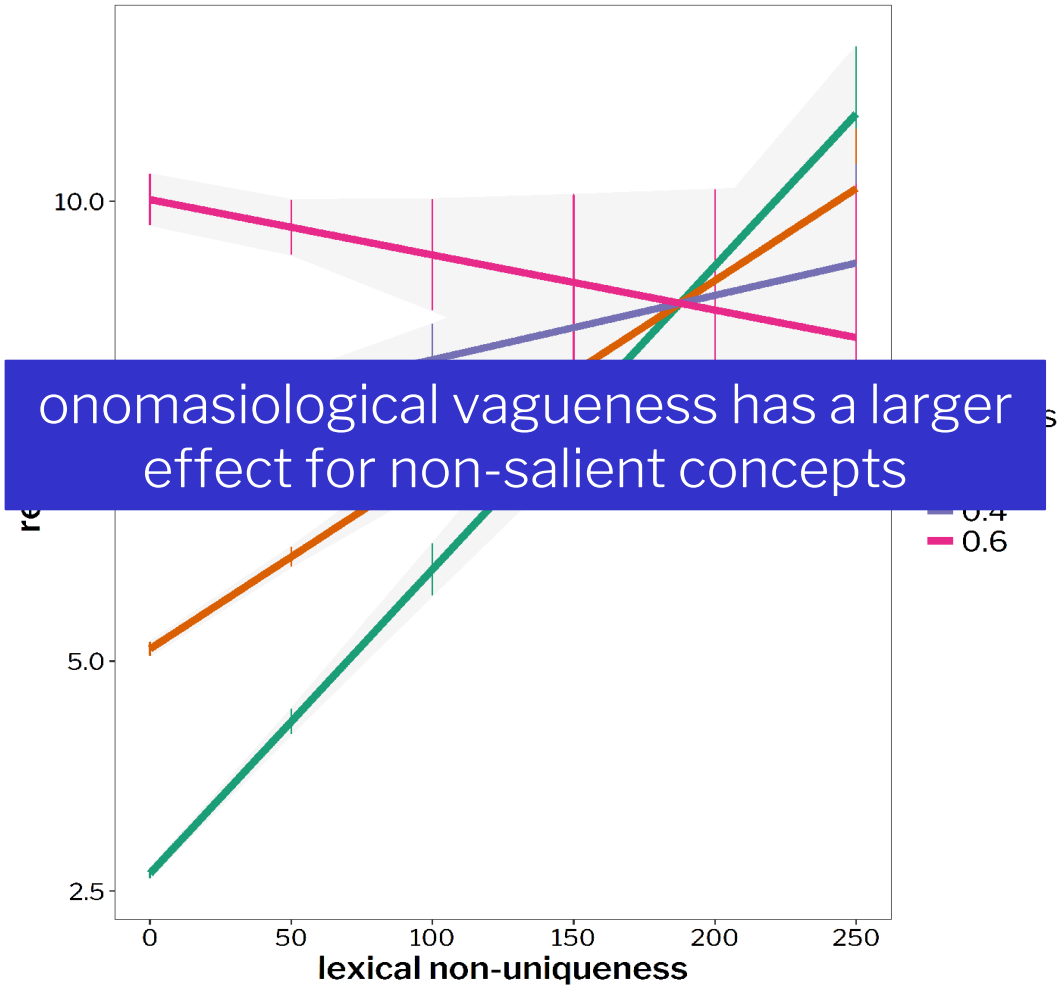
interactions with semantic field



importance of concept characteristics differs per semantic field



lack of onomas. salience * onomas. vagueness



onomasiological vagueness has a larger effect for non-salient concepts

RESPONSE: CALCULATION

number of types per concept

e.g. TO GET MARRIED (TROUWEN): 3 types

trouwen	181
zich binden	1
getrouwd worden	1

CORPULENT WOMAN (GROF GEBOUWDE VROUW): 131 types

machochel	67	mokkel	8
schommel	41	bai (fr.)	7
molenpaard	23	madsel	5
machine	17	schokkel	5
kapitein	11	dikke madam	4
mangel	11		...

geographical fragmentation

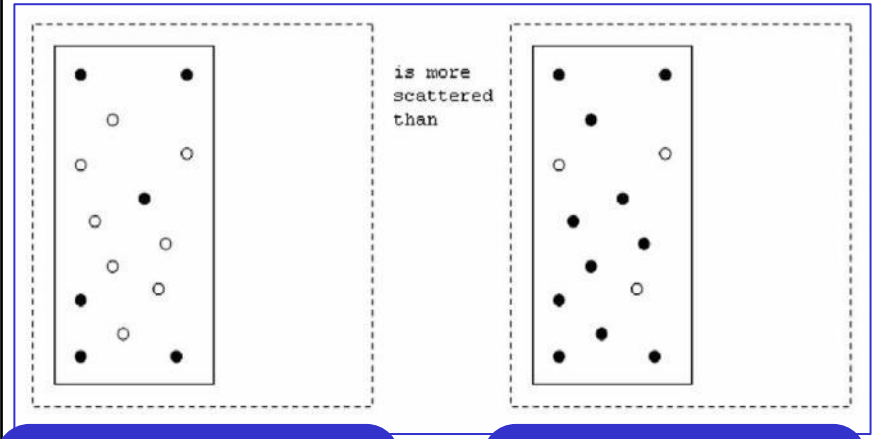
- calculated as the proportion of dispersion and range

- dispersion: (weighted) average distance between the attestations of the unique words for a concept relative to other words for the same concept
- range: (weighted) average coverage of the words for a concept relative to the entire region where the concept occurs

(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)

dispersion & range

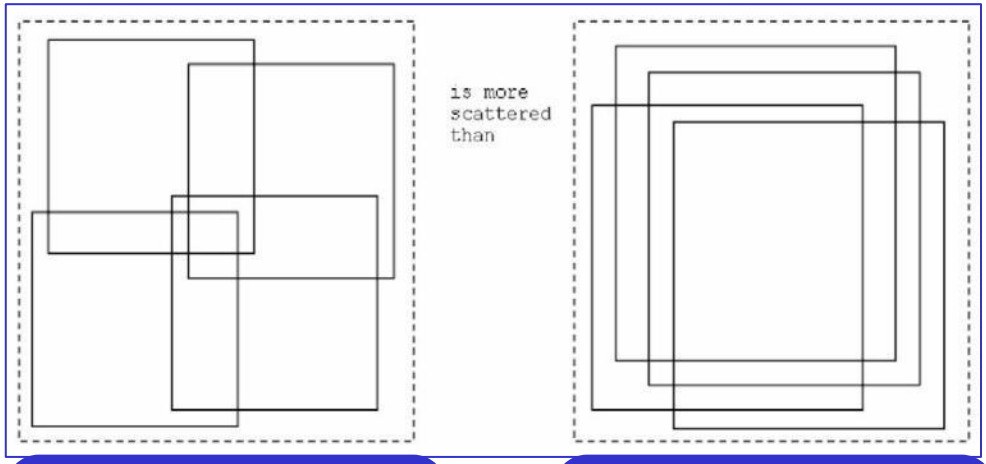
dispersion



variants scattered across dialect area

variants are found in nearby locations

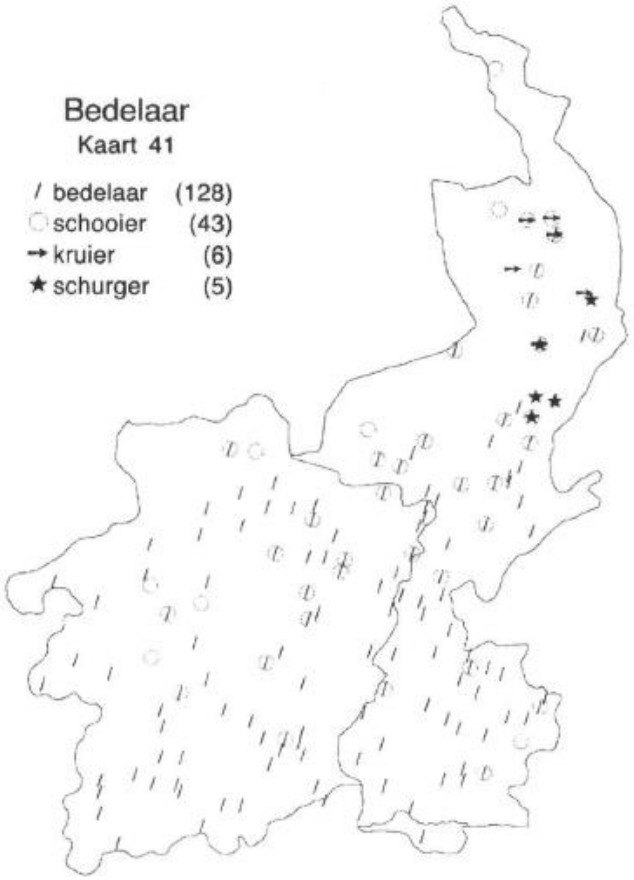
range



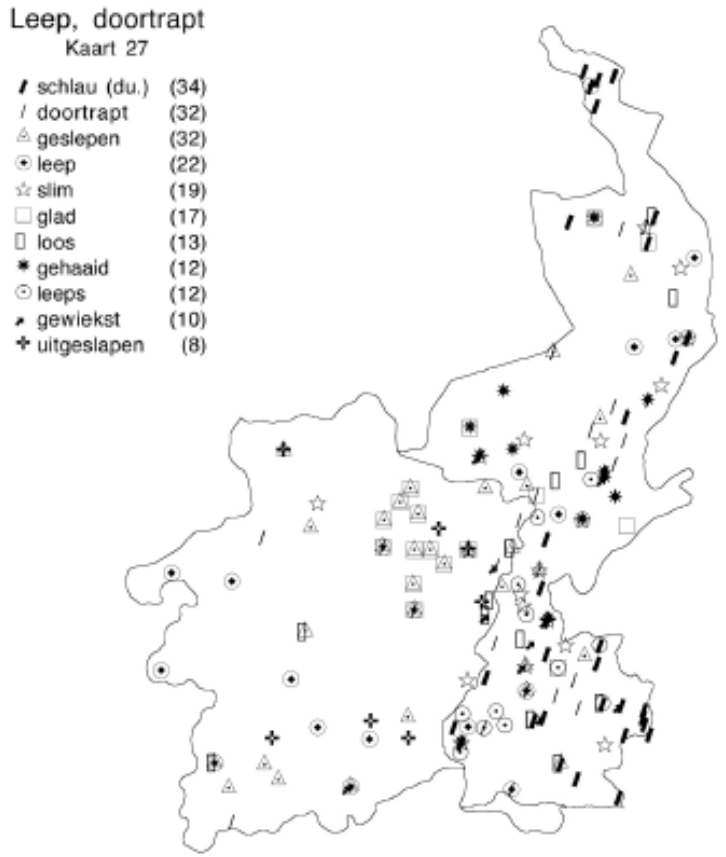
each word type occurs in small geographical area

each word type takes up almost entire dialect area

dispersion



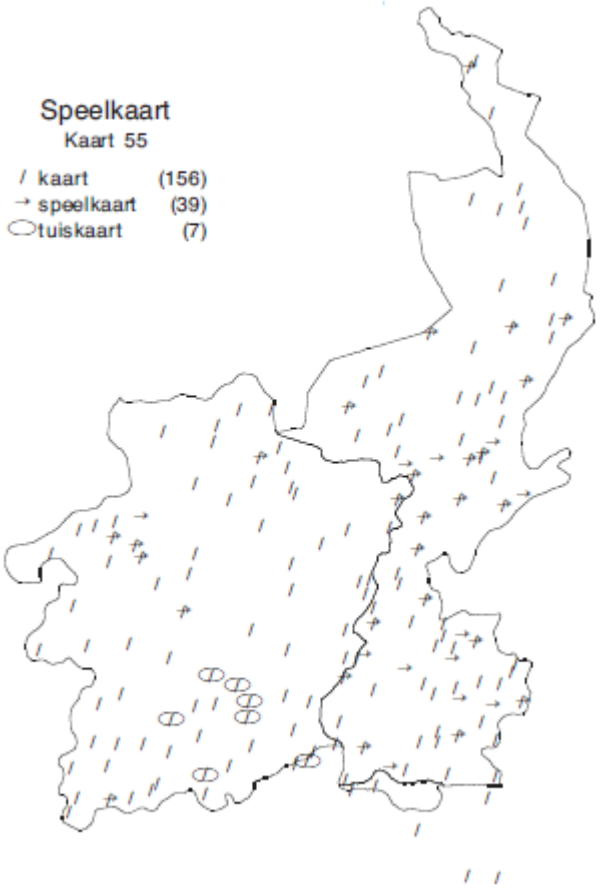
dispersion = 1.22



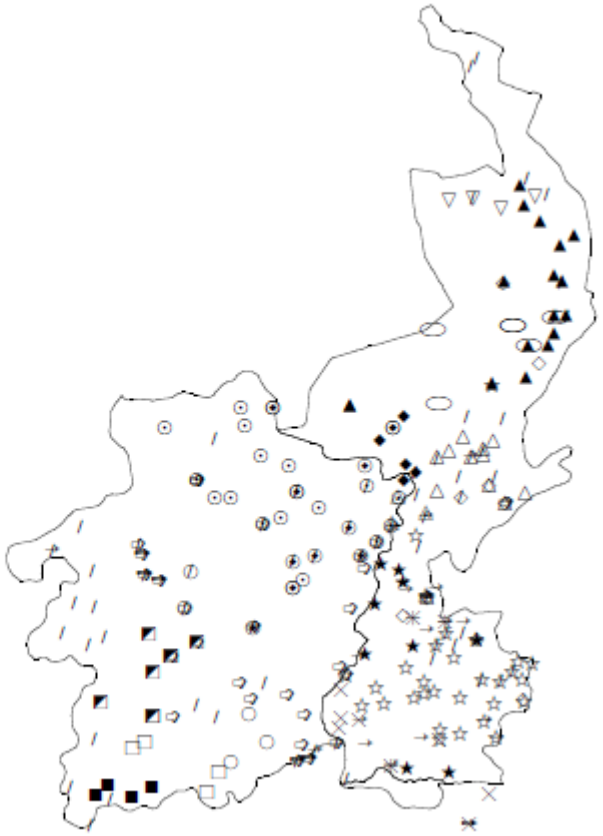
dispersion = 2.58

(between 1 and 4.401)

range



range = 0.82



range = 0.20

(proportion)

GEOGRAPHICAL VARIATION VS. NUMBER OF TYPES

diverging effects: salience vs. vagueness & affect

"the results seem to imply that, while a higher degree of onomasiological vagueness and affect induce both more heterogeneous profiles for the concepts and, over and above geographical fragmentation, a larger amount of unique variants per concept, a higher lack of onomasiological salience only affects the former aspect of lexical diversity."

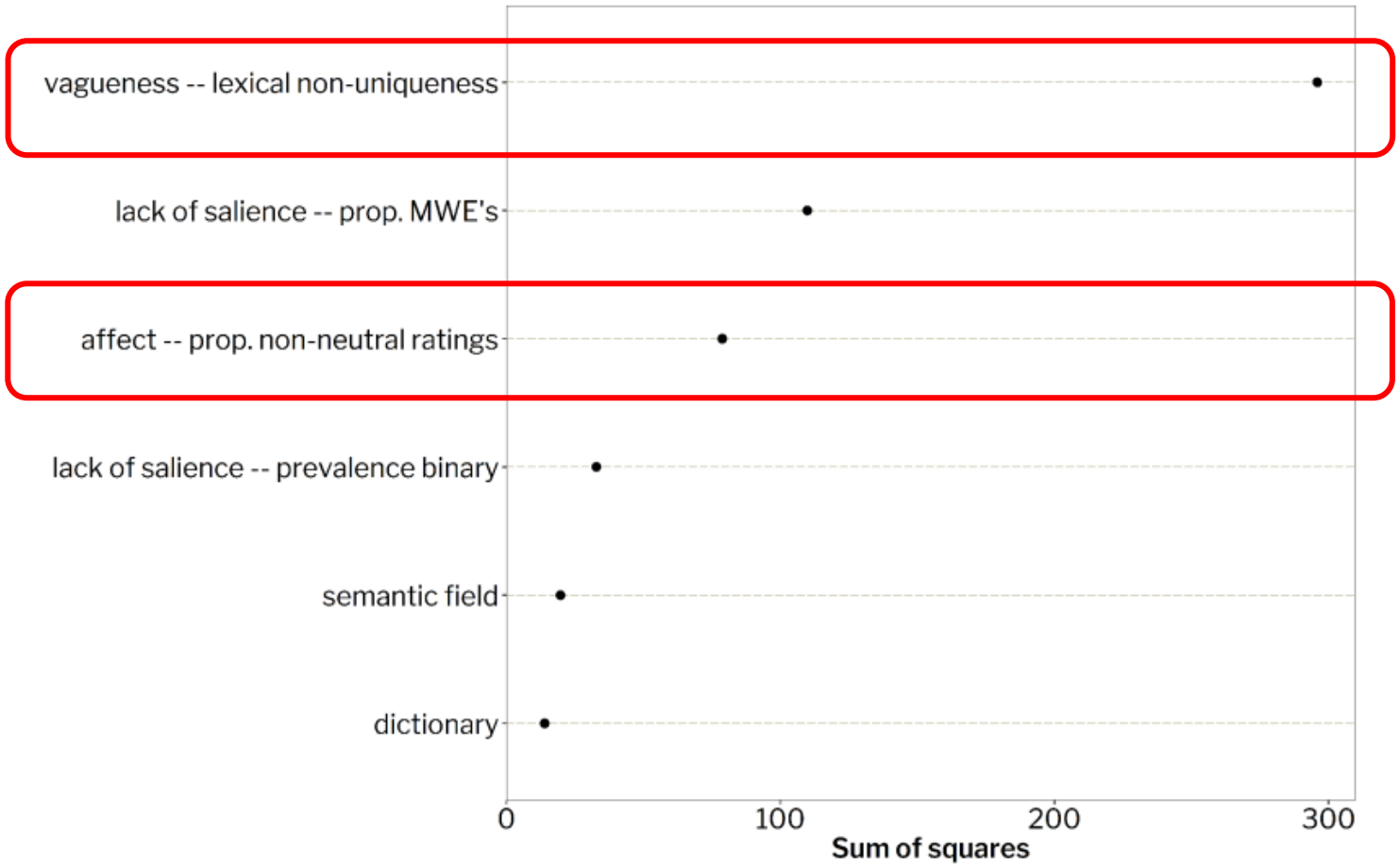
(Franco 2017: 76)

- lack of onomasiological salience: hyperonymous, co-hyponymous, or possibly hyponymous names associated with more salient concepts

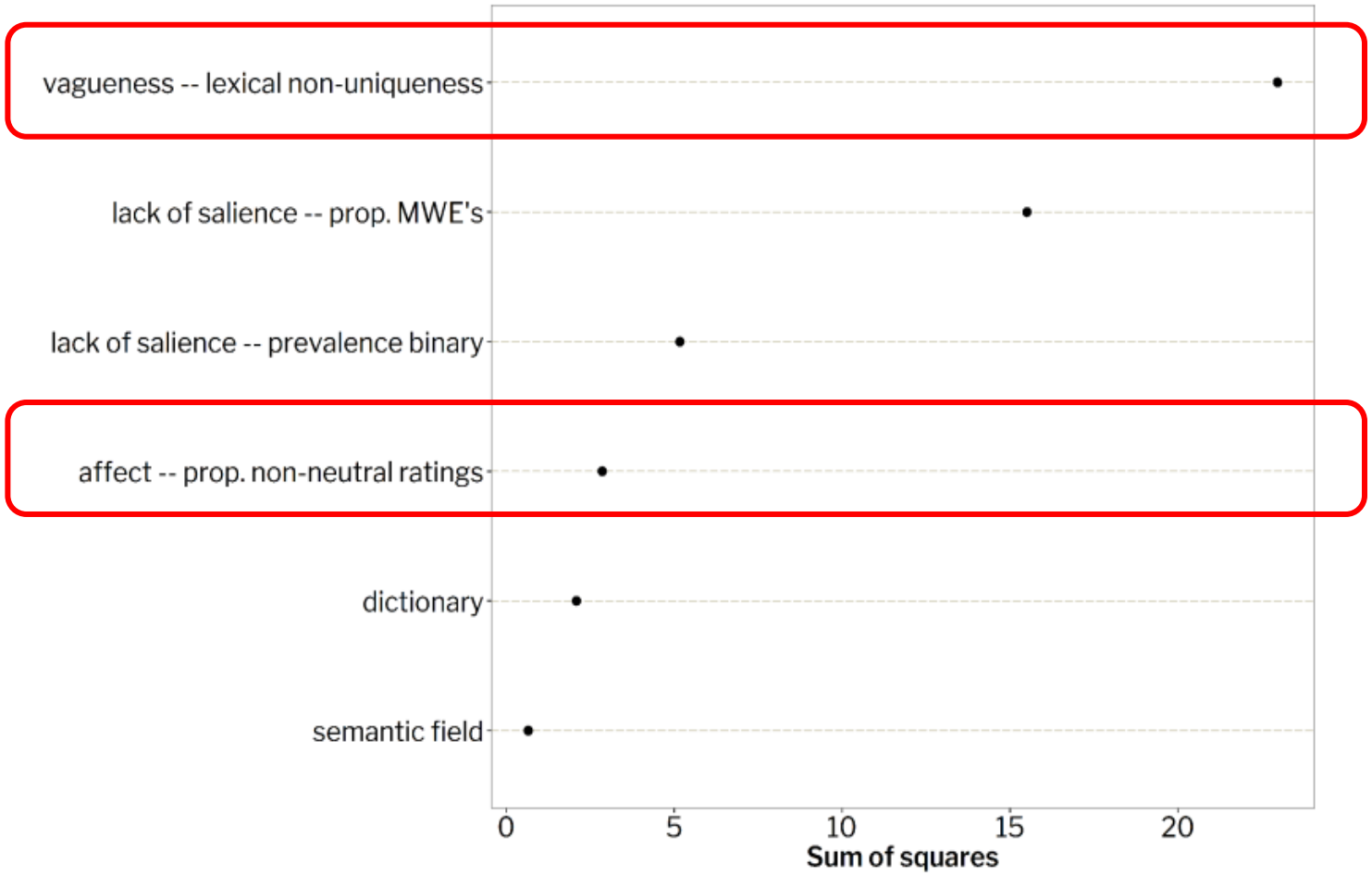
vs.

- affect-sensitive concepts: disposition to lexical creativity
- vague concepts: demarcational differences between speakers

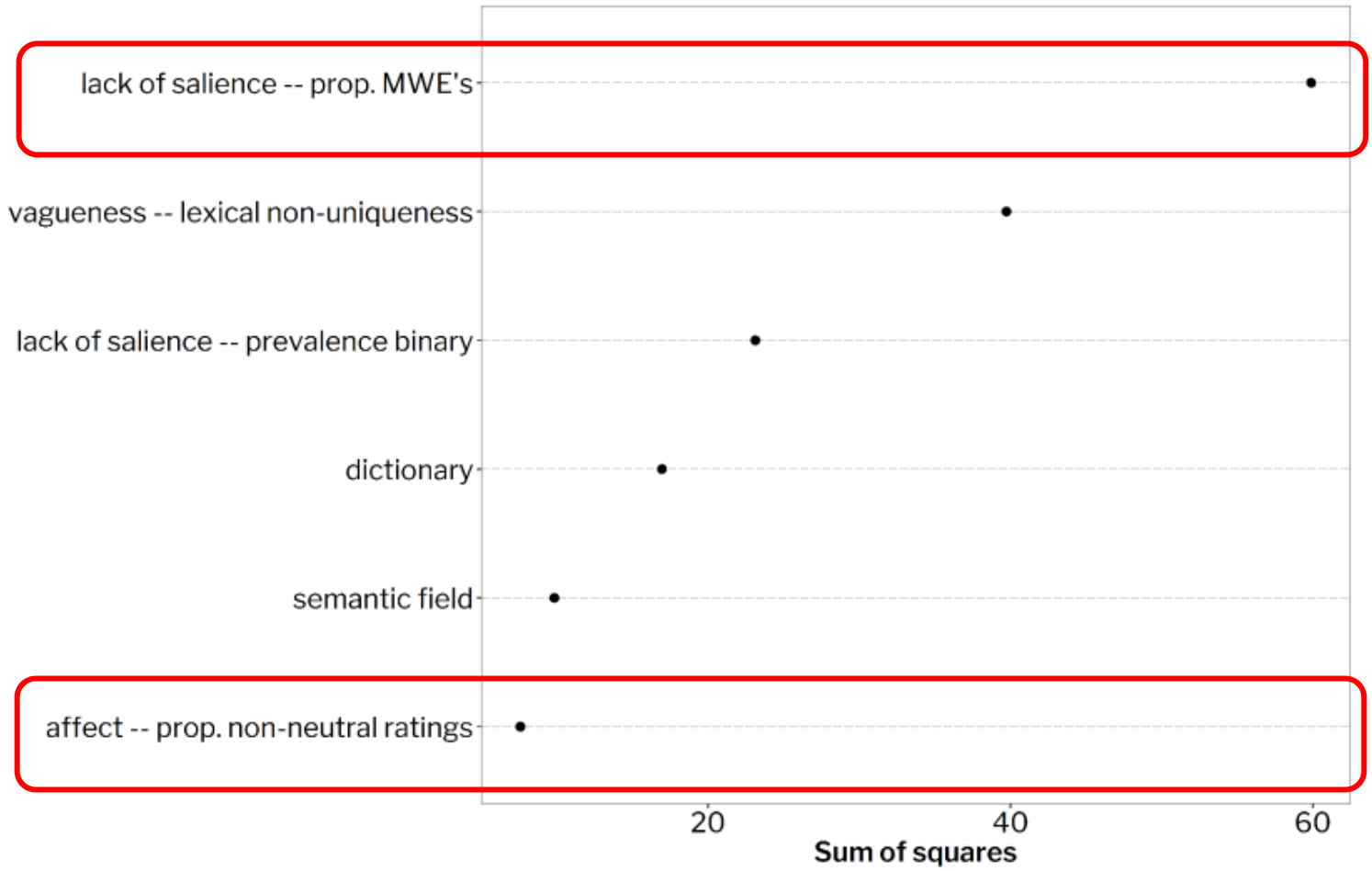
model 1: number of types



model 2: weighted average dispersion



model 3: weighted average range



model term	estimate	SE	p-value
intercept	-0.4420	0.0360	< 0.001
dictionary			
WBD	0.2120	0.0170	< 0.001
semantic field			
the house	0.0130	0.0460	NS
celebration & entertainment	-0.0750	0.0460	NS
personality & feelings	0.2060	0.0510	< 0.001
family & sexuality	-0.0940	0.0700	NS
society, school & education	-0.0730	0.0410	< 0.1
lack of salience			
proportion of hapaxes	2.3460	0.3630	< 0.001
vagueness			
lexical non-uniqueness	0.0220	0.0020	< 0.001
affect			
proportion of non-neutral ratings	0.2070	0.0240	< 0.001
interaction terms			
sem. field (*the house*) : proportion of hapaxes	-1.2440	0.4940	< 0.05
sem. field (*celebration & entertainment*) : proportion of hapaxes	-2.3410	0.4190	< 0.001
sem. field (*personality & feelings*): proportion of hapaxes	-1.8670	0.4040	< 0.001
sem. field (*family & sexuality*) : proportion of hapaxes	1.2250	0.7220	< 0.1
sem. field (*society, school & education*) : proportion of hapaxes	-0.6110	0.4130	NS
sem. field (*the house*) : lexical non-uniqueness	-0.0020	0.0020	NS
sem. field (*celebration & entertainment*) : lexical non-uniqueness	0.0000	0.0030	NS
sem. field (*personality & feelings*): lexical non-uniqueness	-0.0080	0.0020	< 0.001
sem. field (*family & sexuality*) : lexical non-uniqueness	-0.0140	0.0060	< 0.05
sem. field (*society, school & education*) : lexical non-uniqueness	-0.0060	0.0020	< 0.05
proportion of hapaxes : lexical non-uniqueness	-0.0570	0.0040	< 0.001